



Standard
Intelligence

White paper

Departure Rate

The Number That Tells You Where to Look

AUTHOR

Michael Clark
Standard Intelligence

PUBLISHED

July 2026

EDITION

Cross-Sector Paper

WEB

standardintelligence.com

AUDIENCE

For organisations that rely on a human reviewing the output

CONTENTS

Summary	3
1. Every major regime treats the human reviewer as the safeguard	4
2. The research gives little reason to trust that safeguard	5
3. Measurement still has a job: detecting review that leaves no trace	6
4. Defining departure rate	6
5. Reading the number: near zero calls for investigation, not a verdict	7
6. Five companion measures and three conditions make the rate interpretable	8
7. Where departure is costly, the rate measures the interface	10
8. Reporting the measure without overclaiming	10
9. Measuring is safer than not measuring, despite the institutional risk	11
10. Treat human oversight as a system dependency, and rely on it less	12
11. Where review remains, manage it as a performance control	13
Where to start, and what it costs	15
Basis and Standing	16
Sources	17

Summary

Almost every regime that governs automated decisions relies on a human at some point in the process: an overseer who can disregard or override the system, a reviewer of an automated decision, a caseworker exercising a statutory discretion, or a professional making an individual assessment. Each regime expects that involvement to be genuine, yet none says what evidence would show that it is.

The gap falls to the organisation rather than the regulator. Under each of these rules (Article 14 of the AI Act, Article 22 of the GDPR, the public law duty to exercise a discretion and professional duties that mandate individual assessment), whether a review was genuine is a question of fact, and in practice the organisation will be expected to show that it was. An organisation that has told a regulator, a court or a customer that a human reviews the output has made a commitment that it will be asked to evidence.

Departure rate is the first and cheapest evidence an organisation can produce of whether its review leaves any trace. Departure rate is the proportion of decisions in which the person with authority over the outcome disagreed with the system's output and acted to change it: declining what it approved, approving what it declined, or selecting differently from its ranking. For one system, it is computed in a day from logs that, in our experience, most organisations already keep.

A near-zero rate is the reading that demands a response. Across a meaningful population, a reviewer who never exercises their authority to question the output might be confirming an accurate model. They could be working a queue that routes only pre-cleared cases, or genuinely arriving at the same conclusions based on the available facts.

The review could also be a rubber stamp.

Because the number cannot separate these explanations, an organisation relying on the control should be able to say which one applies to it and to show how it determined that answer. If it cannot, a challenger will argue that the review was nominal, and the organisation will have no evidence with which to answer it.

Five companion measures, read against three conditions, make the signal interpretable. The most important is sampled independent review, in which a second competent reviewer, blind to the system's output, decides a sample of cases afresh. The others are time per decision, coverage, recorded reasons for each departure, and abstention and escalation; the conditions are that the reviewer has the competence to judge the decision, the information needed to depart and the authority to act.

The rate should be monitored, never targeted. A minimum departure rate rewards disagreement rather than sound review; a reviewer who overrides correct outputs to show independence improves the metric and worsens the decisions. Organisations should track the distribution, investigate changes against model performance and case mix, and set no quota.

The research points to a harder conclusion: human oversight cannot be relied on to work as policy assumes. Green (2022) surveyed 41 policies that prescribe human oversight of government algorithms and found that they rest on

an uninterrogated assumption that people can oversee algorithmic decision-making effectively. Automation bias affects experts as well as novices, and the literature does not support treating training as a reliable standalone safeguard.

Human oversight is itself a dependency of the system, and the more consequential the system, the less sensible it is to treat a reviewer as an assumed means of correcting its errors. The question to ask of each system is whether it should be relying on a person to catch the error at all. Where it should not, the remedy is to redesign it so that it needs less oversight.

Where review remains, it should be managed as a performance control. Correctness alignment scores each reviewer's overrides against what actually happened, over a horizon that matches the decision: days for a fraud flag, a year or more for a loan. Held-out overrides and an automated-decision control group, used only where the law permits and never by removing a safeguard the law requires, show whether review is making decisions better or only adding cost.

Basis and Standing, at the end, sets out which propositions are binding law, which are regulator guidance, and which are our own recommended controls.

1. Every major regime treats the human reviewer as the safeguard

AI regulation, data protection, public law and professional regulation all assume that a person will catch what the system gets wrong.

The AI Act separates the system, the provider, the deployer and the person overseeing it. High-risk systems must be designed and developed so that they can be effectively overseen (Article 14(1)), a requirement the provider must ensure is met (Article 16), and the provider must supply the system so that the people assigned to oversight are enabled, as appropriate and proportionate, to:

- (a) understand the system's capacities and limitations, and monitor its operation;
- (b) remain aware of the tendency to rely or over-rely on its output (automation bias);
- (c) interpret its output correctly;
- (d) decide, in any particular situation, not to use it, or to disregard, override or reverse its output;
- (e) intervene in its operation or interrupt it, bringing it to a safe halt.

The deployer must then use the system in line with its instructions for use and assign oversight to people with the necessary competence, training and authority, as well as the necessary support (Article 26(1) and (2)). Every item in the list describes what the overseer must be able to do, not what they actually do, and a system can meet all five while the person using it approves every output unread.

Data protection law looks at what actually happens. It prohibits in principle, or makes conditional, decisions based solely on automated processing that have legal or similarly significant effects, and a decision falls outside those rules only where human involvement is meaningful rather than nominal. Guidance from the Article 29 Working

Party, endorsed by the EDPB, says oversight must be “meaningful, rather than just a token gesture” and carried out by “someone who has the authority and competence to change the decision”; UK GDPR Article 22A states the test in terms.

Public law reaches the same position by a different route. A discretion conferred by statute must be genuinely exercised, and a policy applied without willingness to consider departing from it is an unlawful fetter (*British Oxygen Co Ltd v Minister of Technology* [1970] UKHL 4, [1971] AC 610). Professional and sectoral regimes assume that a qualified person makes the decision and can account for it.

Each regime accepts automated support because a human stands between the system and the outcome. None of these regimes says what evidence would show that the human is doing the job, and the organisation’s exposure begins in that gap.

2. The research gives little reason to trust that safeguard

Any measure of oversight has to start from the research on whether oversight works. The weight of that research is that human oversight cannot be relied on to work as policy assumes.

The central critique. Green (2022) surveyed 41 policies that prescribe human oversight of government algorithms and found that they rest on an uninterrogated assumption that people can oversee algorithmic decision-making effectively; reviewing the evidence, he finds that people are largely unable to perform the oversight these policies assume. He concludes that such policies “provide a false sense of security in adopting algorithms”, and proposes a shift to institutional oversight in which agencies must justify an algorithm with evidence; where there is not sufficient evidence that human oversight is effective and that the algorithm improves decisions, governments should not incorporate it into human decision-making.

Automation bias and complacency. Much of the experimental work behind that critique predates the policy debate. Parasuraman and Manzey’s review (2010) reports both effects in naive and expert participants, and finds that where the decision aid is imperfect, automation bias produces errors of omission and commission. Goddard, Roudsari and Wyatt’s systematic review in the medical informatics literature (2012) describes the same pattern, in which people follow incorrect automated advice they would not otherwise have followed, including by abandoning decisions that were correct.

Training is not a reliable fix. The usual organisational response to a weak oversight control is more training. Training and instructions may reduce automation bias, and Goddard, Roudsari and Wyatt list training and user accountability among its mitigators alongside interface design, but the literature does not support treating them as a reliable standalone safeguard.

The findings are not uniform. A public-sector paper reporting three experiments found no automation bias. It did find selective adherence to advice that matched stereotypes

in a citizen sample, whether the advice came from an algorithm or a human expert, but not among civil servants (Alon-Barkat and Busuioc, 2023).

That leaves an organisation with two tasks: finding out whether its own review is doing anything, and relying on review less wherever it can.

3. Measurement still has a job: detecting review that leaves no trace

Green's critique and departure rate, the measure defined in section 4, address different questions, so one does not undermine the other. Green asks whether human oversight prevents harm. Departure rate asks a narrower question, whether the review leaves any trace of independent judgement in the organisation's own records.

The narrower question matters because it targets the failure every regime in section 1 is concerned with, a decision that is reviewed on paper and automated in practice. Organisations rarely detect this failure in themselves, because the workflow records a review and a signature whether or not any judgement was exercised.

Departure rate makes the failure visible. A reviewer who has never departed across a large number of decisions may be confirming an accurate model or exercising no judgement at all. The rate cannot say which, but it shows the organisation that it is relying on a control with no visible effect, and that in practice it will need to explain why.

A near-zero rate is not proof of a breach, and should not be treated as one. What the rate can show depends on how precisely it is defined.

4. Defining departure rate

The measure. Among decisions where a system produced an output and a person had authority over the outcome, departure rate is the proportion in which that person disagreed with the output and acted to change it.

Unit of analysis. Compute the rate per system and per reviewer within each system. The aggregate hides the individual, and the individual distribution is usually where the finding lies; a team with an overall rate of four per cent may contain four reviewers at fifteen per cent and eleven at zero.

Period. Use a window long enough to give volume and short enough to show change, monthly for high-volume processes and quarterly otherwise. Track the rate as a series rather than a snapshot.

What counts as a departure. A departure changes the outcome, not merely the route to it. Adjusting a score without changing the decision is not a departure; reversing a

recommendation is. Where a system ranks rather than decides, a departure is a change in what was selected.

Exclusions. Exclude cases where the system abstained or produced no confident output, since there was nothing to depart from. Report separately the cases routed to a person because the system flagged uncertainty, since the base rate of error there differs by design.

Near zero. Judge it against volume rather than a fixed percentage. No departures across several hundred decisions is a reading to investigate; twenty decisions are too few to yield a rate at all.

Even a well-defined rate is easy to misread, and in our experience most implementations go wrong at interpretation rather than calculation.

5. Reading the number: near zero calls for investigation, not a verdict

A near-zero rate creates an evidential question the organisation should be able to answer. Across enough decisions to be meaningful, a rate at or near zero has several possible explanations:

- an accurate model, where the system is right on this case mix and there is little to correct;
- a filtered queue, where a threshold routes only clear-cut cases to the reviewer and the hard ones never arrive, which is benign for the reviewer but a coverage gap for the organisation (section 6);
- genuine agreement, where the reviewer examines each case and independently reaches the same outcome;
- an easy population, where few decisions are borderline and there is little room to disagree;
- a reviewer not exercising judgement, who confirms the output without examining it.

The number looks the same in every case. A loan reviewer who approves every recommendation from a well-calibrated model and one who approves every recommendation unread both produce a departure rate of zero.

The organisation therefore needs to know which explanation applies and how it knows. Without that, a challenger's argument that the reviewer was not exercising judgement goes unanswered. The right response to a near-zero rate is an investigation rather than a conclusion, using the measures in section 6.

A very high rate signals a different problem. A rate of forty per cent means the people using the system do not accept its outputs, which points to a model that performs poorly, reviewers who have not been trained to use it, or a tool that should not be deployed.

Every report should state the base-rate limitation. The base rate is the system's error rate: how often it is actually wrong. Departure rate measures how often reviewers

respond, and a response rate can only be interpreted against that base rate, which organisations generally do not know.

Signal detection theory, which treats the reviewer as a detector trying to identify inaccurate or unfair outputs (Langer, Baum and Schlicker, 2025), makes the point formally: how often a detector fires means nothing without knowing how often there was something to detect.

Two reviewers who each depart in five per cent of cases show why. If the system is wrong in two per cent of cases, the first reviewer is overriding more often than the system errs, so some overrides must be changing correct outputs. If the system is wrong in twenty per cent of cases, the second reviewer is catching at most a quarter of the errors and letting the rest through.

Departure rate should therefore be reported alongside whatever is known about the system's error rate, even if only from sampling. Where nothing is known, the level of the rate says nothing about the quality of review, and only the extremes and a sustained trend warrant investigation.

Direction matters more than level. A rate that falls steadily over several quarters, with no change to the model or the case mix, indicates reviewers deferring more and more to the system. The fall can show in the series before it shows in complaints, appeals or audit findings, which makes the trend the reading most likely to give early warning.

Monitor the rate; never target it. Any minimum or quota turns disagreement into the goal. Movement in the rate is a prompt to look at model performance and case mix, not a score for the reviewer.

6. Five companion measures and three conditions make the rate interpretable

On its own, departure rate is a weak signal. Two measures from the same logs, time per decision and coverage, make it considerably stronger; three more, which take more effort, are needed before drawing any conclusion.

Time per decision, as a distribution. This is how long the reviewer spends on each case before confirming or changing the output, and its shape tells more than its average. Very short and highly compressed review times may be a warning signal, particularly where they do not vary with case complexity, but they do not by themselves establish whether meaningful review occurred. A competent reviewer may assess a simple case in seconds, and a long review does not prove a meaningful one.

Coverage. Coverage is the share of decisions that received any human attention. A process that reviews only the cases the system escalates has checked the system's conclusions on those cases, not how the system behaves overall, and the cases it never surfaces are often the ones that matter. A recruitment screen that passes ten shortlisted candidates from four hundred applicants to a hiring manager has put ten decisions in

front of a person; the three hundred and ninety rejections were made by the system alone.

Read together, the three log measures answer separate questions. Coverage shows whether anyone looked, time shows how long they looked, and departure shows whether looking changed anything. They show where review may be nominal, but they do not prove it.

Sampled independent review. This measure carries the most weight. A second competent reviewer takes a sample of decisions the first reviewer approved and decides them afresh, without seeing the system's output. If the two largely agree, that is evidence the first reviewer's agreement with the system reflects a correct model; if they often differ, it suggests outputs were being confirmed without examination.

The measure is only as good as the second reviewer, who must be competent and genuinely unable to see what the system decided. It is also the measure that takes the most effort, which is why it is usually skipped. None of the others can substitute for it, because only this one tests whether agreement with the system was right.

Recorded reasons on departure. When a reviewer overrides the system, the organisation should check that a reason is recorded and that it engages with what the system produced. "Declined: does not meet criteria" says nothing, while "declined: the model's income figure omits a second job shown on the payslips" shows that the reviewer examined the case. A departure without that kind of reason tells the organisation as little as no departure at all.

Abstention and escalation. This measures how often the reviewer declines to decide, or passes the case to someone more senior or more specialised. A process with no route to abstain forces a decision in every case. A claims handler who cannot refer an unusual medical claim to a clinician must accept or reject it, whether or not they are able to judge it.

Three conditions sit alongside the measures. The reviewer must have the competence to judge the decision, the information needed to depart and the authority to act. Where any one is missing, a low departure rate says more about the conditions than about the reviewer.

Competence comes first, because a reviewer who cannot assess the case on its merits has no basis for disagreeing with the system. An administrator asked to confirm a model's fraud flags, with no training in what fraud looks like, will accept them because there is nothing else they can do. Article 26(2) of the AI Act, set out in section 1, makes this competence a legal requirement for high-risk systems.

Competence is necessary but not sufficient. It gives the reviewer a basis to disagree, but section 2 shows that training, the usual way to build it, is not a reliable standalone safeguard against automation bias.

Information means seeing enough of the case, and of the basis for the output, to form a view. Section 7 deals with it alongside the cost of departing.

Authority means the power to act on a departure without escalation or personal consequence. Where overriding carries a cost that agreeing does not, reviewers will override less, and section 7 shows how that cost is built into the design.

7. Where departure is costly, the rate measures the interface

The information and authority conditions determine whether the rate means anything. Where departing from the system is costly, the rate measures that cost rather than the reviewer's judgement.

Take a reviewer who can accept a recommendation with one click but must write a justification, raise a variance report and wait for a manager's sign-off to override it. If overriding takes fifteen minutes where accepting takes two, or draws attention that agreement does not, a low rate is a rational response to the design.

The first step is to find out whether departing is costly. Before reading a departure rate, sit with a reviewer, watch them override a decision and count the steps, screens and people involved. If departing is hard, the number is measuring the interface.

The fix sits in the interface too, which is usually the cheapest place to make it. Making departure as easy as acceptance is a design decision, for example an override that takes the same number of steps as approval and asks for a reason rather than a sign-off. So is showing the reviewer enough of the basis for an output to disagree with it, since a reviewer shown only a score, with none of the inputs behind it, has nothing to argue with.

The research on further interface changes is encouraging but mixed. Cognitive forcing functions are designs that interrupt the reviewer and require engagement with the case before the system's output is shown, for example by asking for the reviewer's own answer first. Buğınca, Malaya and Gajos (2021) found that they reduced overreliance, but users rated the most effective designs least favourably and the benefit was greater for people with a higher need for cognition, meaning those who enjoy effortful thinking.

Explanation design, which shows the reviewer why the system reached its output, has a similar record. Neither is a solved problem, so an organisation adopting either should treat it as an intervention to be measured, by watching whether departure rate and the results of sampled independent review change, rather than a control to be assumed.

8. Reporting the measure without overclaiming

How an organisation describes the results matters as much as how it calculates them. The temptation is to overstate, and an overstatement is what a challenger will attack.

What an organisation may say. It may say what it measured and what it did about the results: that it measures departure rate, time per decision and coverage for each system, alongside sampled independent review and whatever is known of the error rate; that the values sit within stated ranges and the series is stable or improving; and that it

investigated any near-zero readings, with what it found. A defensible statement might read: “The departure rate on one of our credit decisioning systems fell to near zero in the second quarter; we traced it to a routing threshold that sent reviewers only clear cases, and changed it.”

What it may not say. On these measures alone, it may not say that oversight is effective, that decisions are meaningfully reviewed or that human judgement is being exercised. None of these follows from the numbers. Asserting them turns a defensible measurement practice into a claim that one case can disprove, such as a single file in which a reviewer approved an obviously wrong output.

Once outcomes are measured. Where correctness alignment (section 11) has run over a full horizon, an organisation may go further and say how often its reviewers’ overrides were right, and therefore what review adds to the system’s decisions. That is the strongest claim available, and the first that speaks to whether review works.

For a board report. Present the three log measures as a series for each system, with the error rate and the results of sampled independent review, name the systems at or near zero and say what was done about each. Leave out any summary line declaring that oversight is operating effectively; the board needs to see where review may be nominal, not a reassurance the numbers cannot support.

For a regulator. The strongest position has three parts: the organisation measures what the regime cares about, acts on adverse readings and claims no more than the measurement supports. That position can be evidenced, which makes it more defensible than a policy that asserts meaningful oversight with nothing behind it.

9. Measuring is safer than not measuring, despite the institutional risk

The recommendation to measure is rarely rejected on its merits. More often it is resisted for an unstated reason: computing the number creates evidence that can be used against the organisation. If the departure rate comes back at zero across a year of decisions, the organisation has produced, at its own expense, the central exhibit in any future claim that its human oversight was nominal.

Not measuring can therefore look like the safer choice. Three considerations point the other way.

The number already exists. Departure rate, time per decision and coverage all sit in the case management system, the workflow logs and the audit trail. Declining to compute them does not stop others computing them; it means the first person to do so is likely to be someone with an adverse interest, working from disclosure, when the organisation can no longer act on the answer.

Ignorance is not a defence. An organisation that asserts meaningful human oversight while holding data that would test the assertion, and never testing it, is likely to be worse

placed than one that measured, found a problem and fixed it. The second has a record of remediation; the first has an unexamined claim and must explain why it never examined it.

Early discovery keeps remediation open. A near-zero rate found today and traced to the review design can typically be fixed within one or two quarters, through reviewer capacity, the three conditions in section 6 and the interface changes in section 7. The same number found in three years describes a period that cannot be revisited, covering every decision made in it, and it can do damage each time it is disclosed, compared with earlier representations or used in a claim.

For a board, measuring is not free of risk, but it turns an open-ended, unmanaged exposure into a bounded, dated problem with a remediation plan. Not measuring leaves the number in existence and cedes control over who computes it first, and when.

10. Treat human oversight as a system dependency, and rely on it less

Section 2 shows that human oversight cannot be relied on to work as policy assumes. A reviewer is therefore a dependency of the system like any other component, and the more consequential the system, the less sensible it is to treat that reviewer as an assumed means of correcting its errors. The question to ask of each system is whether it should be relying on a person to catch the error at all; where it should not, there are three ways to place less weight on review.

Fewer decisions that rely on review. Where the system's output decides the outcome and the human step is nominal, the decision is in substance automated. If the law permits a solely automated decision of that kind, the organisation should describe it as one and meet the obligations that follow, rather than keep a review that satisfies nobody. Where applicable law in effect requires meaningful human involvement (GDPR Article 22 where none of its exceptions applies, UK GDPR Article 22B where it applies, or public law wherever a body must exercise its own discretion), the organisation should either design the human stage so that it satisfies that requirement or reconsider the use and role of the system in the decision.

The AI Act does not itself prohibit solely automated decisions by high-risk systems. Where review of each decision is the oversight measure the provider specifies or the deployer adopts (Articles 14(3) and 26), however, a nominal review will not meet it.

Tighter limits on what systems may do unsupervised. Constraining what a system may do without a person is more reliable than assuming a person will catch what it does. Three constraints work without anyone having to notice anything in the moment: suspending the system when its measured performance falls below a set level, having it abstain on inputs unlike the data it was built on, and setting hard limits on consequential actions, such as a ceiling on the value of payments it can release.

Controls that do not depend on catching errors in real time. Evidence captured at the point of decision, and monitoring that detects drift across all cases rather than one at a time, both hold up without relying on a reviewer's vigilance.

11. Where review remains, manage it as a performance control

Departure rate shows whether reviewers act. It does not show whether acting makes decisions better, and that is the question oversight exists to answer. Model monitoring already answers the equivalent question for the system, by comparing each prediction with what actually happened, and the same discipline should apply to the people overseeing it.

We call this correctness alignment: measuring each reviewed decision against its eventual outcome, over the period in which that outcome becomes known, so that the human layer is held to the same standard of correctness as the system. Run continuously, it is the ongoing performance management control for oversight.

The horizon should match the decision. A fraud flag can be confirmed or cleared within days. A lending decision is right or wrong only when the loan is repaid or defaults, which may take a year or more, and a hiring decision only once the person's performance can be assessed.

For a predictive model, the horizon is the window the model is judged on. A credit model built to predict default over twelve months should have its reviewers' overrides judged over the same twelve months; judging them sooner measures something other than correctness.

Outcomes are observed only for the decision actually taken. If a reviewer declines a loan the model would have approved, the organisation never learns whether that applicant would have repaid, so the override cannot be scored. Credit risk practice knows this as the reject inference problem, and for oversight it means that overrides which withhold something cannot be judged from outcomes alone.

Overrides are not a random sample. Reviewers choose which cases to override, and they may choose the hardest or most ambiguous ones. As a result, observed outcomes are not sufficient to estimate the causal value of review without an appropriate counterfactual design. Two designs can supply one, held-out overrides and an automated-decision control group, and both raise legal questions that come before any statistical ones.

A control group that removes a legally required human safeguard is not a permissible experiment merely because it is statistically useful. Any counterfactual design means some people receive a decision that a reviewer, or the organisation's usual process, would not have given them. Where the decisions have legal or similarly significant effects, such a design is likely to involve decisions based solely

on automated processing, including where a reviewer's override is set aside so that the system's decision stands.

Under the GDPR such decisions are permitted only within an Article 22(2) exception and with its safeguards, and under UK GDPR they must meet Article 22B and carry the Article 22C safeguards, including the right to obtain human intervention.

The AI Act adds two constraints for high-risk systems. A held-out override sits uneasily with the ability the provider must give the overseer to override the output "in any particular situation" (Article 14(4)(d)) and with the authority the deployer must assign (Article 26(2)), and a control group must stay within the oversight measures in the instructions for use, which the deployer must follow (Article 26(1)). Where a public body must exercise its discretion in every case, neither design may be possible.

Where testing is lawful, the two designs work as follows:

- **Held-out overrides.** Where a reviewer withholds something the system would have granted, such as declining a loan the model approved, the system's decision stands for a small random share of those overrides, and the outcome is recorded against the reviewer's judgement. Overrides in the other direction need no hold-out, because the outcome of what the reviewer granted is observed anyway.
- **An automated-decision control group.** A small random sample of cases is decided by the system alone, without review, and its outcomes are compared with those of reviewed cases of the same type. This shows whether review as a whole improves decisions or only adds cost and delay.

Both should be as small as the statistics allow, confined to decisions whose harm is bounded and reversible, and designed with legal advice. A held-out override suspends the reviewer's authority for that case, so reviewers should be told that hold-outs are in use, though not which of their cases are held out. Where neither design is lawful or acceptable, sampled independent review (section 6) is the best available substitute.

The result is a score for each reviewer and each system. Override accuracy is the share of a reviewer's overrides that the outcome later showed to be right, measured through the designs above, and read alongside the system's accuracy on the cases the reviewer left alone. Together they supply the error rate that section 5 says departure rate cannot be interpreted without.

Read per reviewer, over enough cases to be meaningful, the results show who is improving decisions. A reviewer whose overrides are mostly right is adding value, one whose overrides are no better than the system's is adding cost without benefit, and one who never overrides produces no score, which is itself a finding.

Like departure rate, override accuracy should be monitored, never targeted; it should inform training, competence records and authority. A reviewer judged on it alone learns to override only when certain, and the departure rate falls. For the systems that must still rely on a person to catch the error, this is how an organisation knows whether the person is right.

Where to start, and what it costs

The recommendations come in stages, and the first costs almost nothing.

The diagnostic: a day, one analyst, no budget. Departure rate, the time-per-decision distribution and coverage can be produced for one system from existing logs. It is a query rather than a project, and it shows which systems have review that leaves no trace.

The remedy, if needed: a small team over one or two quarters. Where the diagnostic finds a rate at or near zero, the work is an investigation using sampled independent review and recorded reasons, followed by work on the three conditions in section 6, interface changes that make the basis of the output visible, and capacity modelling. Where review cannot be made real, the remedy is the redesign in section 10.

The ongoing control: for as long as the system is in use. Correctness alignment scores reviewers' overrides against outcomes over the horizon that matches the decision. Its cost is the work of linking decisions to outcomes, plus any held-out overrides or control group the law permits, and it is what turns oversight from an asserted control into a measured one.

On timing. The AI Act's oversight provisions apply from 2 December 2027 for Annex III high-risk systems and 2 August 2028 for those under Annex I. Under Article 111(2) as amended, systems placed on the market or put into service before then are caught only after a significant change in design, or by 2 August 2030 if intended for use by public authorities. As at 28 July 2026, no fine under the AI Act had been confirmed by any regulator or official source.

The exposure described here, however, is not primarily an AI Act exposure. It arises wherever an organisation has told a regulator, a court or a customer that a human reviews the output, and in our experience most organisations have already made that statement in writing, where it can be tested.

That statement is a commitment the organisation will be asked to evidence. Departure rate is where the evidence starts and correctness alignment is how it is kept current; an organisation that runs both can say not only that a human reviews the output but what that review adds and, where it adds nothing, can stop relying on it.

This paper describes an approach to measurement and is not legal advice. The research cited on the efficacy of human oversight is contested and developing, and the regulatory requirements described are stated as at 28 July 2026. Take advice on the obligations applicable to your own systems.

Basis and Standing

Binding law. AI Act Article 14 on human oversight as a requirement of high-risk systems, which providers must ensure is met (Article 16), and Article 26(1) and (2) on deployers' use of the system and assignment of oversight to persons with the necessary competence, training, authority and support, applying from 2 December 2027 (Annex III) and 2 August 2028 (Annex I) under Article 113 as amended by Regulation (EU) 2026/1744. Under Article 11(2) as amended, high-risk systems placed on the market or put into service before those dates are caught only if their design changes significantly from then, and those intended for use by public authorities must comply by 2 August 2030. EU GDPR Article 22, and UK GDPR Articles 22A to 22D (substituted for Article 22 by section 80 of the Data (Use and Access) Act 2025, in force 5 February 2026), on solely automated decisions; Article 22A defines solely automated decisions by the absence of meaningful human involvement. Public law prohibits fettering a statutory discretion (*British Oxygen Co Ltd v Minister of Technology* [1970] UKHL 4, [1971] AC 610).

Sector requirements vary and are named where used.

Regulator guidance. Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling (WP251rev.01, endorsed by the EDPB), on when human involvement in a decision is meaningful, including that it be carried out by someone with the authority and competence to change the decision.

Not a regulatory metric. Departure rate is not defined, prescribed or endorsed by any regulator. It is this paper's proposed diagnostic, offered because it is computable from data organisations already hold. Coverage, time per decision, sampled independent review, recorded reasons, and abstention and escalation in section 6 are likewise this paper's design, as are correctness alignment, override accuracy, held-out overrides and the automated-decision control group in section 11; the competence and authority conditions reflect Article 26(2) and WP251, and the information condition reflects the support Article 26(2) requires and the abilities Article 14(4)(a) and (c) require the system to enable.

Research, and what it does and does not support. The literature cited largely supports the proposition that human oversight of algorithmic decisions cannot be relied on to work as policy assumes, including work on automation bias, complacency and overreliance. It is not uniform. Alon-Barkat and Busuioc (2023) found no evidence of automation bias across three experiments; one found selective adherence to advice that matched stereotypes, whether it came from an algorithm or a human expert, and a study of civil servants did not. The literature does not validate departure rate as a measure of meaningful oversight, and no claim is made that it does. The inference limits are set out in the summary and section 5.

Our experience. The statements made in our experience, that most organisations already keep the logs the rate needs, that most implementations go wrong at interpretation and that most organisations have already stated in writing that a human reviews the output, are practitioner observation rather than findings.

Do not set a target. A minimum or target departure rate creates a gaming incentive and is expressly not recommended.

Enforcement position. As at 28 July 2026, no fine under the AI Act had been confirmed by any regulator or official source.

Sources

1. AI Act, Articles 14, 16, 26, 111 and 113. <https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng>
2. GDPR Article 22. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>; Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling, WP251rev.01 (endorsed by the EDPB), p.21. <https://ec.europa.eu/newsroom/article29/items/612053>; UK GDPR Articles 22A to 22D. <https://www.legislation.gov.uk/eur/2016/679/article/22A>
3. Ben Green, “The Flaws of Policies Requiring Human Oversight of Government Algorithms”, *Computer Law & Security Review*, vol. 45, 105681 (2022), surveying 41 human oversight policies. <https://arxiv.org/abs/2109.05067>; DOI <https://doi.org/10.1016/j.clsr.2022.105681>
4. Raja Parasuraman and Dietrich H. Manzey, “Complacency and Bias in Human Use of Automation: An Attentional Integration”, *Human Factors*, vol. 52(3) (2010), pp. 381-410 (subscription access). <https://journals.sagepub.com/doi/10.1177/0018720810376055>
5. Saar Alon-Barkat and Madalina Busuioc, “Human-AI Interactions in Public Sector Decision Making: Automation Bias and Selective Adherence to Algorithmic Advice”, *Journal of Public Administration Research and Theory*, vol. 33(1) (2023), pp. 153-169 (subscription access). <https://academic.oup.com/jpart/article/33/1/153/6524536>
6. Kate Goddard, Abdul Roudsari and Jeremy C. Wyatt, “Automation bias: a systematic review of frequency, effect mediators, and mitigators”, *JAMIA* vol. 19(1) (2012), pp. 121-127. <https://academic.oup.com/jamia/article/19/1/121/732254>
7. Z. Buçinca, M. B. Malaya and K. Z. Gajos, “To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making”, *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (2021). <https://doi.org/10.1145/3449287>
8. *British Oxygen Co Ltd v Minister of Technology* [1970] UKHL 4, [1971] AC 610. <https://www.bailii.org/uk/cases/UKHL/1970/4.html>
9. Markus Langer, Kevin Baum and Nadine Schlicker, “Effective Human Oversight of AI-Based Systems: A Signal Detection Perspective on the Detection of Inaccurate and Unfair Outputs”, *Minds and Machines*, vol. 35, article 1 (2025). <https://doi.org/10.1007/s11023-024-09701-0>

Provisions are cited by article, rule or paragraph in the text above.



Standard
Intelligence

GET IN TOUCH

Find out whether your human review leaves a trace.

EMAIL

hello@standardintelligence.com

WEB

standardintelligence.com

LOCATION

Manchester, United Kingdom

AI Use: AI tools were used in research, editorial review and drafting support. All sources were retrieved and checked against the published record by the author, who holds editorial responsibility for the contents and for the conclusions drawn. This disclosure is made voluntarily. To the extent the Article 50(4) obligation in Regulation (EU) 2024/1689 applies to this paper as text published to inform the public on matters of public interest, the exception for content that has undergone human review and for which a person holds editorial responsibility applies.