Human Oversight Is Not a Presence: It Is a Capacity Under Constraint

Two doctrines, one word

The French doctrine of the *garantie humaine* (human guarantee) and the human oversight of the AI Act do not designate the same object, and public debate has been conflating them for five years. This confusion has a practical consequence: we continue to require human oversight without ever measuring whether it is one.

The first doctrine crystallised around article L. 4001-3 of the French Public Health Code, created by article 17 of Law no. 2021-1017 of 2 August 2021. That text provides for the information of the person concerned, the information of the health professionals resorting to the processing, access to the data used and to the results derived from them, and the explicability of the system's functioning, incumbent upon its designers. It does not itself employ the expression *garantie humaine*. That expression belongs to the doctrine that has surrounded the text, from CCNE opinion no. 130 of May 2019 to the joint opinion CCNE 141 and CNPEN 4 of November 2022, which recognises human guarantee colleges bringing together professionals and patient representatives. This doctrine is multidisciplinary and deferred: it organises a collective gaze upon a device, upstream and downstream of its use.

The human oversight of article 14 of Regulation (EU) 2024/1689 belongs to another logic. It requires that high-risk systems be designed, human–machine interfaces included, in such a way that natural persons can effectively oversee them during the period in which they are in use. The regulation does not require that this oversight be exercised decision by decision, nor in real time. But in the class of systems that concerns us here (serial decision support at sustained cadence) the requirement becomes principally individual and synchronous: it must be exercisable at the workstation, within the useful time of the decision.

The first doctrine was promised while the second device was funded. This text bears on the second, and on that class of systems alone. It does not hold for rare decisions with high unitary stakes, where deliberation time exists by construction.

Saturation, not volume

Article 14, paragraph 4, enumerates what the operator must be able to do. These capacities can be sorted into three families. To act, first: to disregard the system's output, to refuse to use it, to interrupt its operation. To understand, next: to know the capacities and limits of the system, to interpret correctly what it produces, to monitor its operation. And a third, of a different nature from the first two: to remain aware of one's own tendency

to rely excessively on the output produced. The first family belongs to ergonomics, the second to training, the third to metacognition.

None of this is a property of the system. These are capacities exercised by a person, at a workstation, under a load.

The title of this note says *throughput*. It is a working metaphor, assumed as such: it installs at the outset an engineering intuition where the regulatory vocabulary offers only moral qualities. It is not the fundamental variable, and that must be said immediately. Volume is not oversight throughput, and oversight throughput is not saturation. What saturates a workstation is not the flow of decisions but the product of that flow by the quantity of judgement each decision demands, related to the cognitive capacity actually available. Two hundred files a day of which one per cent present a real ambiguity may be perfectly controllable. Twenty files a day each requiring a quarter of an hour of independent analysis may not be. The same workstation, the same person, two opposite regimes.

$$S = \frac{\lambda \times J}{C}$$

where λ is the frequency of decisions submitted to the workstation, J the average judgement demand per decision, and C the effective judgement capacity available. Three regimes follow from it.

1. When $S \ll 1$, the workstation has a cognitive reserve and oversight can be what it claims to be.
2. When $S \to 1$, it becomes fragile without anything signalling it.
3. When $S \geq 1$, systematic independent review is not degraded: it is structurally impossible, and what subsists under that name is a ratification procedure.

This ratio is a design heuristic proposed here, not a calibrated instrument borrowed from cognitive ergonomics. Neither J nor C is directly measurable, and writing it this way does not make them measurable. The ratio is moreover written on average values. In operation, its distribution matters still more: a workstation that is sustainable on average may saturate under the effect of clusters of cases with high judgement demand, and it is precisely at those moments that oversight is most necessary. What the notation brings is a displacement of the question: to stop asking how many files the operator processes, and to begin asking how many judgements the workstation demands and how many it can produce.

It also brings a consequence that the literary formulation concealed. The introduction of an AI system does not act on a single term, it acts on all three, and not in the same direction. It increases λ, since that is generally the object of the deployment. It reduces J on simple cases, which it handles well. It increases J on residual cases, which are precisely those it handles badly and leaves to the human. And it reduces C over the long

run, through deskilling. A device may therefore degrade its own oversight without any of its indicators moving.

When this ratio approaches unity, oversight does not disappear: it changes in nature. It becomes a procedure of ratification or of default rejection, whose output no longer depends on the content of the case but on the configuration of the workstation.

Under load, oversight converges towards the action that the workstation makes free. An alert is an interruption, the free action is to dismiss it, load produces mass rejection. A review workstation is a validation station, the free action is to approve, load produces ratification. Same mechanism, two settings of the default. The design variable is therefore not vigilance, it is the setting of the default action and the distance separating the operator from saturation.

Four conditions, and they multiply

Effective oversight presupposes four conditions: a judgement capacity, an epistemic independence, an effective authority, and the sustainability of disagreement. They do not add up. In the heuristic proposed here, they behave multiplicatively,

$$\boldsymbol{H = f(S) \times I \times A \times D,\ f'(S) < 0}$$

where H denotes the effective oversight capacity and f a decreasing function of the saturation defined above. The form of f is not specified, and cannot be: nothing today allows us to say whether the degradation is gradual, threshold-based or abrupt. The hierarchy of the two quantities matters more than their form. S is a property of the workstation under load. H is a functional property of the oversight device installed at that workstation. A value close to zero on a single term suffices to annul the contribution of the other three. This is a property of the model, not a measured regularity.

Judgement capacity, the first condition, is what saturation consumes: it is the term $f(S)$, and it is not read off the number of files processed.

Epistemic independence. This is probably the most poorly understood condition, because it is reduced to information. It would be excessive to maintain that judging an output always requires information the system has not consumed: an operator may detect an error from the same data, through a different representation, method or mental model. The independence required may be informational, methodological, temporal, organisational or causal. The real problem is not the sharing of data, it is **the correlation of errors**. If the human and the system exploit the same data, rely on the same intermediate variables and have been trained by the same institutional practices, adding the human produces an apparent redundancy without a proportionate gain in safety. Safety engineering has known this for a long time under the name of common-mode failure, and the vocabulary of successive barriers employed in this note is that of James

Reason (*Human Error*, Cambridge University Press, 1990; *Managing the Risks of Organizational Accidents*, Ashgate, 1997).

This point far exceeds article 14, and that is what makes it useful. Redundancy is not independence. **Two readers trained in the same school do not necessarily constitute two independent barriers.** Two models sharing their training data and their architecture do not either. And the limiting case deserves to be named, because it is becoming the implementation norm: an operator who verifies an output on the basis of the explanation supplied by the system that produced it is not a second barrier, it is a **pseudo-redundancy**. The explanation improves understanding and simultaneously frames the reasoning of the one who receives it; its provenance therefore forms part of the architecture of the barrier, on the same footing as its content. **The value of a second barrier does not depend on its own performance, but on the correlation of its errors with those of the first.** Two decision-makers do not make two barriers when their failure modes are correlated.

Authority. The regulation names it, but only where it truly designs something: recital 48 and article 14, paragraph 5, require that the persons assigned to oversight have the necessary competence, training and authority. Authority is not the button. An operator may have the time, detect the error, have at hand a perfectly ergonomic override command, and exercise no effective oversight if their disagreement triggers a hierarchical justification, degrades their indicators, or may be held against them later. I have proposed elsewhere to name this gap between an actor's formal authority and their real grip on the decision. It constitutes here the third condition, and it is the only one the regulation mentions explicitly, in a provision it applies to a single case.

The sustainability of disagreement. Distinct from authority, with which it is readily confused. Authority is the permission to deviate. Sustainability is what deviating consumes, and how many times a day one can consume it. In the ordinary regime, ratifying consumes little while deviating commits, documents, slows down and is seen. The asymmetry is not a law of nature: it reverses as soon as an organisation demands accounts for ratifications as much as for overrides, which remains rare. We shall see later that when this asymmetry does reverse, it is generally too late and before a commission of inquiry.

Here must be named the most serious counter-thesis, the one a lawyer will raise first. Article 14 does not impose a cognitive result. It requires the provider to make oversight possible by design, and the deployer to entrust it to persons having the necessary competence, training and authority. To reproach the regulation for not guaranteeing vigilance amounts to reproaching it for not being what it never claimed to be. The objection is correct in legal terms, and that is precisely where the institutional problem lies: an obligation of designed means, uninstrumented, produces a documentary

compliance whose effectiveness no one, including after the fact, can establish. The text is defensible. The use expected of it is not.

It will also be objected that the saturation of alerting devices owes less to volume than to low specificity, hence to false positives. The objection is correct, and it refines the thesis rather than defeating it: improving specificity reduces the judgement demand per decision, that is to say acts on the numerator of the saturation ratio. It is a confirmation of the model, not a refutation.

A remark on the state of the law remains. Regulation (EU) 2026/1744, which entered into force on 27 July 2026, rewrote article 4: providers and deployers must take measures to support the development of AI literacy among their staff, without that obligation requiring them to guarantee a determined level for anyone. It would be easy, and wrong, to see in this a contradiction with article 14. The latter remains an autonomous obligation, and the relaxation of article 4 dispenses no one from organising effective oversight where it is required. The tension lies elsewhere, and it is more worrying than the contradiction: article 14 defines functional capacities that the oversight device must make it possible to exercise, foremost among them the understanding of the system's limits and resistance to automation bias, at the very moment when article 4 weakens one of the horizontal means liable to contribute to them. The design requirement remains owed. The means has become optional in its intensity.

Aviation, read correctly

The two sections that follow change argumentative regime. They proceed not by measurement but by authority and by case: what the human factors literature has established, and what one documented affair shows. The data return in section 6.

The aviation analogy circulates in both directions, and generally badly.

It does not prove that the human is the weak link. Statistics attributing the majority of accidents to a human cause must be handled with caution: the very category of human error partly reflects the manner in which socio-technical systems attribute causality and responsibility. Madeleine Clare Elish described this mechanism under the name of the moral crumple zone, where the operator closest at hand absorbs a responsibility even though their effective control was limited, which protects the integrity of the technical system (*Engaging Science, Technology and Society*, 2019, vol. 5, pp. 40–60, DOI 10.17351/ests2019.260). It is a serious methodological objection, not a general refutation of the statistics.

What aviation does establish is more useful to our purpose. Lisanne Bainbridge formulated it in 1983 in *Ironies of Automation* (*Automatica*, vol. 19): automation leaves to the human the tasks he performs least well — prolonged monitoring and recovery in abnormal situations — while degrading, through disuse, the very skills he will need on the

day automation fails. The connection with the saturation ratio is direct, and it constitutes its dynamic reading. Bainbridge describes exactly what the introduction of a system produces on the three terms: the judgement demand JJ concentrates on the rare and difficult cases, those automation does not handle, while capacity CC erodes for lack of exercise on ordinary cases. The numerator increases where it is most costly, the denominator decreases in silence. This is not a historical reference, it is the transient regime of the model.

The sector that invests the most in human vigilance has been documenting its degradation for forty years. This is an a fortiori argument, not a transposition: constrained traffic, selected operators, institutionalised feedback — nothing comparable with a workstation processing scored files on a production line.

And aviation did not obtain its safety by demanding vigilance. It ceased to depend on it: flight envelope, redundancy, enforceable checklists, crew coordination, dual signature, recorders, non-punitive feedback. This reading commits the author and is not here supported by a history of aviation safety. It is nonetheless verifiable on the record: in that sector, human oversight is an instrumented and recorded process, not a moral quality expected of the operator.

In the whole regulation, a single place proceeds this way. Article 14, paragraph 5, imposes verification by at least two natural persons for remote biometric identification. One case, not a method.

What unmeasured oversight becomes

The Dutch childcare benefits affair is the best-documented case, and one must take care not to make it carry a single model. Amnesty International, in *Xenophobic Machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal* (2021), describes there a scoring system, the use of nationality as a risk factor, a manual review required after flagging, the absence of information enabling agents to understand the score, and the absence of meaningful human oversight. The decisional architecture is that of article 14: the machine flags, the human examines. But the causality there is richer than the load effect alone: discrimination embedded in the variables, opacity of the output, rigidity of administrative rules, institutional asymmetry, recovery incentives. Nominal human oversight coexisted with an opaque score and an environment that prevented manual review from constituting an independent barrier. Our model explains part of the mechanism. It does not explain the affair.

It does, on the other hand, illuminate what becomes of the asymmetry of cost. For years, ratifying a flag cost the agents nothing. Then the government's resignation in January 2021 made those ratifications extremely costly, retrospectively, and for the persons least well placed to contest them. This is the moral crumple zone at the scale of an administration.

Ben Green reviewed forty-one policies imposing human oversight and draws two conclusions from them (*The Flaws of Policies Requiring Human Oversight of Government Algorithms*, *Computer Law & Security Review*, 2022, vol. 45, DOI 10.1016/j.clsr.2022.105681). The first is the empirical incapacity of persons to fulfil the expected functions. The second, more disturbing, is that these policies legitimise the adoption of defective systems by producing a false sense of security. Nominal oversight does not protect the person exposed to the decision. It protects the deployment, and it exposes the operator.

Human oversight is a classifier, therefore it is evaluated

If nominal oversight is the problem, measurement is the answer. Still, the right thing must be measured, and the obvious measure measures nothing.

Two systematic reviews have established the variability of override rates for medication alerts in computerised prescribing. Van der Sijs and co-authors, across seventeen publications, report override rates from 49% to 96% (*Journal of the American Medical Informatics Association*, 2006, PMID 16357358). A more recent review covering twenty-three studies published between 2000 and 2019 reports a range from 46.2% to 96.2%, with marked variation by alert type, from 46% to 95% for drug allergy alerts and from 82% to 96.8% for dosing alerts (*JMIR Medical Informatics*, 2020, PMID 32706721).

At the other end of the spectrum, Rosbach and co-authors measured the inverse movement in computational pathology: in an online experiment bringing together twenty-eight trained pathologists, tasked with estimating a percentage of tumour cells, exposure to an erroneous recommendation leads to the reversal of an initially correct assessment in approximately 7% of cases (arXiv:2411.00998, 2024). Time pressure did not increase the frequency of the phenomenon but aggravated its severity. What this study proves: automation bias exists among experts, on a task at which they are competent, and it is quantifiable. What it does not prove: a prevalence in routine clinical practice, the sample being twenty-eight and the protocol experimental. Of note, and the text would be dishonest to omit it: in the same study, the integration of AI significantly improved overall performance.

These two series of figures do not add up. The first pertains to under-reliance, the second to over-reliance. What they establish together is that a high override rate does not prove effective oversight, and that a low rate does not prove a reliable system.

The JMIR review goes further, moreover, and this is its most useful result here: the authors classified the overrides according to their appropriateness, and obtain a range running from 29.4% to 100% depending on the alert type. In other words, two departments displaying the same override rate may, in one case, exercise pertinent oversight and, in

the other, produce noise. The rate does not separate them. The qualification of each override does.

The correct formulation is therefore this one, and it must be bounded in order to hold. Article 14 attributes several functions to oversight: to interpret, to contextualise, to refuse to use the system, to interrupt it, to detect anomalies. Only one of them lends itself to the treatment that follows. When one attributes to human oversight the function of detecting and correcting erroneous outputs, it becomes itself a classifier of those errors. And a classifier is evaluated. It takes as input a machine output and produces a binary verdict, ratification or override. A classifier is evaluated by cross-tabulation with the true state.

	AI incorrect	**AI correct**
Human overrides	useful override	unnecessary override
Human ratifies	error let through	correct ratification

From this table are derived the four quantities that really describe an oversight device: the sensitivity of the oversight, its specificity, the predictive value of disagreement and the predictive value of ratification. The evaluation framework is the banal one of any diagnostic test. Its application to the device of human oversight is, by contrast, a proposal of this note, and it is its core: it moves human oversight from the status of a disposition to be documented to that of a function whose performance is measured.

This requirement extends a position I defend elsewhere on evidentiary requirements: an organisation either records its evidence or reconstitutes it, and what it has not recorded, it will not demonstrate. Human oversight does not escape the rule. It is even its most revealing case, since it is the only safety device whose presence is demanded without the log ever being asked for.

A precaution is then called for, and it stands as a warning against naïve use of the table. An output error is not an incorrect decision, and an incorrect decision is not a harm. A false output may remain without consequence, a technically exact output may contribute to an unjust decision, a correct override may arrive too late, a faulty ratification may be caught by a barrier situated downstream. Three layers of performance must therefore be distinguished: that of the system, that of the oversight, that of the socio-technical device as a whole. The sensitivity of the oversight measures the second. It does not measure the final reduction of risk, and no one should present it as such. This is also the reason why human oversight cannot be evaluated alone: in an architecture of barriers, the performance of one barrier says nothing about the safety of the whole so long as one is ignorant of what subsists behind it.

An oversight whose sensitivity is unknown is not weak. It is not evaluated.

There remains the serious difficulty, and it is real: in most high-risk systems, the true-state column is not available at the moment of the decision. In credit, in recruitment, in social protection, in justice, the error is observable with delay, counterfactual, or contestable. To demand measurement without addressing this point would be a hollow injunction precisely where it is most necessary.

Responses exist, and they are known to quality systems: deferred ground truth when the outcome becomes observable, independent double reading on a sample, expert adjudication, reference cases injected periodically into the flow, systematic review of disagreements, counterfactual analysis when the protocol permits. None is free. All are less costly than the after-the-fact demonstration that a displayed oversight was not one.

This requirement is not merely good practice. In the SCHUFA Holding judgment of 7 December 2023 (C-634/21), the Court of Justice did not ask who signed, but what determined: a score constitutes in itself an automated individual decision as soon as the third party receiving it draws on it in a determining manner. The Court ruled nothing on nominal human review — the question was not put to it, and it would be imprudent to ascribe that solution to it. But it recalls at paragraph 67 that the data controller must be in a position to demonstrate its compliance. An organisation that does not know the sensitivity of its oversight does not possess, by that formal organisation of oversight alone, a measure enabling it to establish in what proportion the machine output is effectively corrected when it ought to be.

The draft European standard that is to specify what human oversight means technically, prEN 18229-3:2026, prepared by committee CEN/CLC/JTC 21, has been under enquiry since this summer. What is at stake in that document is whether article 14 will be operationalised by measurable requirements or by a list of dispositions to be documented.

Instrument, invert, declare, or move up a level

There exists a documented device that holds, and it acts on the saturation ratio rather than on vigilance. The MASAI trial, conducted on 105,934 participants in Sweden, compared AI-assisted mammographic reading with standard double reading. The system directs examinations scored 1 to 9 towards a single reading and those scored 10 towards a double reading. The publications report a 44.2% reduction in reading workload (Lång et al., *Lancet Oncology*, 2023, 24(8), pp. 936–944, DOI 10.1016/S1470-2045(23)00298-X), a 29% increase in detection without a significant rise in false positives (Hernström et al., *Lancet Digital Health*, 2025, 7(3), e175–e183, DOI 10.1016/S2589-7500(24)00267-X),

and, on the primary endpoint, a non-inferior interval cancer rate, 1.55 versus 1.76 per thousand, ratio 0.88 with a 95% confidence interval from 0.65 to 1.18 (Gommers et al., *The Lancet*, 2026, 407(10527), pp. 505–514, DOI 10.1016/S0140-6736(25)02464-X).

What these figures prove: at a reading workload reduced by 44%, the device does not degrade the result and detects more. What they do not prove: a superiority on interval cancers, the confidence interval covering unity. The protocol moreover holds within an organised screening programme, a perceptual task and expert readers. The human there remains fully in the interpretive loop, and it would be false to present MASAI as an absence of oversight. The innovation lies elsewhere: the device does not ask a constant quantity of human attention to follow an increased flow, it uses the machine to allocate that scarce resource differentially. In terms of saturation, it acts first on the numerator, by reducing the judgement demand where it is not necessary. It perhaps also preserves the denominator, by concentrating the readers' exercise on the cases that merit it, but the documented effect is the fall in workload.

Whence two levels of response, which must not be placed on the same plane.

At the transactional level, three patterns. To measure, that is to say to know the sensitivity of the oversight and to make it an auditable asset. To reduce the judgement demand, using the machine to bring JJ back to the level at which judgement exists. Or to design the system in such a way that the human does not have to compensate for a structural weakness, which amounts to declaring the oversight device for what it is and shifting the safety burden upstream.

At the institutional level, something else. Green does not conclude in favour of better human oversight, he concludes in favour of displacing the centre of gravity towards institutional oversight. This means ceasing to ask the operator to catch every error, and controlling instead what really falls to the organisation: the admission of the system, the populations concerned, the thresholds, the conditions of use, post-deployment monitoring, the conditions of suspension and withdrawal.

Govern → Admit → Operate → Oversee → Monitor → Suspend

Transactional oversight occupies one place in this sequence, and only one. It is a barrier of last resort at the level of the decision, it is not the safety architecture itself. Treating it as such is precisely the error that the regulation encourages without prescribing it.

Limits

Four, at least.

1. The class of systems targeted is narrow. The reasoning does not apply to rare decisions with high unitary stakes, where the judgement resource is not the active constraint.

2. The empirical corpus mobilised on override rates comes overwhelmingly from healthcare, often from rule-based alerting systems, and not from Annex III systems. It holds as an argument of mechanism, not as a transposable measure. The three authors who structure sections 4 and 5 (Bainbridge, Elish and Green) belong moreover to the same critical tradition of human factors, which orients the framing.
3. The saturation ratio is a heuristic, not a calibrated instrument. Its two terms are approached through indirect indicators whose validity remains to be established.
4. Finally, the proposed measurement has a feedback effect. Making the override rate visible is to make it an object of steering, hence of optimisation. A device evaluated on its disagreement rate will produce disagreement. This is a known risk of any control metric, and it is treated only by cross-tabulation with the true state, that is to say by the whole table and not by its first row.

A fifth limit must be added, and it is the most serious. Ground truth is not always ontological. In imaging, an interval cancer ends up existing or not. In recruitment, in credit, in social protection, the right decision is normative, multidimensional and often counterfactual: there exists no true state to oppose to the verdict, only a second judgement that will itself have to be legitimated. The diagnostic table then remains a useful evaluation framework, but ceases to be a directly measurable matrix. The threshold between the two situations is not technical, it is political, and this note does not settle it.

Conclusion

Human oversight does not exist because a human is in the loop. It exists if that loop possesses a judgement capacity, an independence, an effective authority and a sustainability of disagreement sufficient to modify the system's trajectory, and if the organisation can demonstrate that it does so when the system errs.

Formulated in the vocabulary of safety, this yields a requirement harder than that of the regulation. Let MM be an error of the system, HH the failure of oversight to correct it, and BB the failure of the barriers situated downstream. The probability that a failure traverses all three is written

$$\boldsymbol{P(M \cap H \cap B) = P(M) \times P(H \mid M) \times P(B \mid MH)}$$

This writing presupposes no independence: it is the exact opposite. It renders the dependencies visible and requires that they be estimated. The performance of oversight must be measured conditionally on the errors the system actually produces, and not on the set of all cases. That of the downstream barriers must be measured conditionally on

the errors that have already crossed the preceding barriers. The correlation of failure modes therefore does not invalidate the calculation. It forbids one thing only, and it is common practice: replacing these conditional probabilities with marginal rates measured separately, then multiplying them. Human oversight is not a property of compliance. It is a barrier whose conditional probability of failure must be known.

The pertinent regulatory question is therefore not whether there is a human, but what risk-reduction function is attributed to that human, what their probability of failure is under real operating conditions, and what independent barriers subsist when they fail. Posing the question in this way avoids making human oversight the new regulatory deity after having dismantled the old one.

Human presence is not a safety barrier. It becomes one only if its risk-reduction function is defined, if its failure modes are measurable, and if its failures are not correlated with those of the system it is supposed to control.

An oversight that never produces disagreement is not necessarily an oversight. An oversight whose sensitivity is unknown is not even evaluated.