Human oversight is not external to the system

The system transforms the one who oversees it. On minimum independent human capacity, and on the absence of any instrument to measure it.

1. The paradox

On 8 September 2026, the Wall Street Journal ran an exclusive on the resignation of Jacob Coxon, a pre-training researcher who had passed through OpenAI and then Anthropic, and who reproached these laboratories for racing towards a self-improving superintelligence while gambling with our lives (Ramkumar A., "Anthropic Researcher Quits Over 'Out-of-Control' AI Fears", WSJ, 8 September 2026). He told the newspaper that the world was heading towards the most aggressive of these scenarios, in which, as early as the end of next year, things could be out of control, and he wrote elsewhere that those who build these systems believe they could kill us all before the end of the decade.

This warning must be taken seriously, and what it contains must be noted. Every one of his assertions is indexed to a future: the end of next year, the end of the decade. None bears on a state measured today, and this is not a failure of rigour on his part, it is the nature of the object. A risk whose probability remains undetermined may perfectly well justify precautionary work — uncertainty reduction, containment, robustness, adversarial exercises; it offers, by contrast, little purchase for a quantitative comparison of priorities. A potentially unbounded impact does not ground a ranking, for want of a comparator. There exists another risk, less spectacular, already observable, and directly attached to the regulatory apparatus now in force.

Here it is. The law requires human oversight of high-risk artificial intelligence systems. In so doing, it implicitly assumes that use of the system does not transform the person charged with overseeing it. That assumption is false, and it is the thesis of this article: **human oversight is not an external capacity applied to the system, it is a state variable coupled to it. Horizontal AI governance today defines neither an explicit metric for this variable, nor a standardised protocol for measuring its depreciation and its restoration.**

One must be precise about the nature of the reproach, because its broad version would be false. Some sectors already possess partial mechanisms that serve in its stead: proficiency testing, periodic competence assessment, double reading, inter-observer

agreement, clinical quality control, re-certification. What is missing is not all measurement, it is the measurement of the right thing in the right place. These arrangements were built to monitor variability between practitioners, not the drift of a practitioner exposed to algorithmic assistance. And the horizontal layer, the one that applies to all high-risk systems whatever the domain, contains none of them.

Likewise, to say that the law treats human oversight as a parameter is not an assertion about the letter of the regulation, which obviously contains no formal model. It is an assertion about its implicit ontology. The law treats it functionally as a parameter: it verifies that the capacity is present and never models its evolution under the effect of using the system. The distinction matters, because the gap this text seeks to establish is instrumental and not normative.

2. What we know, what we do not know, and why

The state of the evidence reads at four levels, and they must be held apart. Conflating them is precisely what allows each camp to cite whichever study suits it.

First level: the signal.

In August 2025, The Lancet Gastroenterology and Hepatology published a study its authors qualify as an incidental finding (Budzyń K., Romańczyk M., Kitala D. et al., Lancet Gastroenterol Hepatol 2025;10(10):896-903). Four Polish centres engaged in the ACCEPT trial had deployed a polyp detection aid at the end of 2021, examinations then being allocated between assisted and non-assisted colonoscopies according to date. Between 8 September 2021 and 9 March 2022, 1,443 patients underwent a non-assisted colonoscopy, 795 before the introduction of the tool and 648 after. The adenoma detection rate fell from 28.4% to 22.4%, an absolute difference of six points (95% CI −10.5 to −1.6; p = 0.0089), exposure to AI emerging as an independent factor in multivariate analysis with an odds ratio of 0.69 (0.53-0.89). Nineteen endoscopists, all experienced. A correction has since been published: the indication for the colonoscopy had been omitted from the covariates, which modifies the coefficients associated with the period following the introduction of AI and with age above sixty, without the authors and the editor drawing from it any change of interpretation.

What makes this study notable is not its figure but its object. The literature on assistance measures the performance of the human-machine pair; this one measures what becomes of human performance when the assistance is withdrawn. The shift of estimand is the real contribution, and it is an estimand that evaluation protocols almost never include.

Second level: the attempted refutation.

A prospective multicentre registry-based pragmatic trial, published in 2026 in Endoscopy (Pedersen et al., Endoscopy 2026;58(9):1003-1014), measured the same phenomenon with a distinctly superior protocol: thirteen endoscopists, seven non-experienced and six experienced, 5,013 colonoscopies, and three successive phases — before exposure to the system, during its use, then after its withdrawal. Primary endpoint: the proportion of colonoscopies detecting at least one polyp of at least five millimetres.

Among the non-experienced, this rate rises from 31.9% (216 of 678) in the initial phase to 39.5% (257 of 651) under assistance, adjusted odds ratio 1.43 (1.11-1.84). After withdrawal it settles at 36.3% (173 of 476), and the difference from the initial phase is no longer significant, odds ratio 1.03 (0.79-1.34). Among the six experienced operators, no significant effect of assistance on the endpoint, in any of the three phases. The authors read the absence of post-exposure reduction as an argument against skill erosion.

This point is decisive and must be stated without comfort. The contradiction persists when the comparison is restricted to experienced operators alone. It cannot therefore be imputed to the level of experience. Nineteen Polish experts degrade, six Scandinavian experts do not budge. The candidate explanations are numerous and none is settled: the endpoint is not the same, adenoma detection versus detection of polyps of at least five millimetres; nor is the design, retrospective before-and-after versus prospective in three phases with explicit withdrawal; the power of the experienced subgroup is weak, six operators; the structure and duration of exposure differ; so do the case-mix and the organisational context of the centres; the adjustments retained are not identical; and the very definition of deskilling is not the same from one protocol to the other. The authors of the prospective trial note, moreover, that their experienced group was appreciably older than the non-experienced group, and suggest that a conservatism bias in human-machine interaction could be at work.

Third level: the state of the field.

A systematic review published in July 2026 queried PubMed, Scopus, Web of Science and Google Scholar, identified 11,269 references and retained 29 studies under PRISMA 2020 (Al-Anezi F. M., "Generative Artificial Intelligence in Healthcare: Automation Bias, Deskilling, and Cognitive Implications", J Healthc Leadersh 2026;18:590498). Automation bias appears there in ten studies, the deskilling concern in nine, impairment of diagnostic reasoning in nine as well. The authors describe a literature that is predominantly observational, experimental on simulation, or purely conceptual, and too heterogeneous for quantitative synthesis. In other words, the concern is massively shared and the evidentiary apparatus is not.

Fourth level: the one that is missing.

Establishing that no longitudinal measurement of the rate of depreciation and the rate of restoration exists presupposes a dedicated and reproducible negative search, with databases queried, search strings, cut-off date, inclusion criteria and screening log. It has not been conducted here. Until it is, this text asserts only that such measurements have not been identified, which is not the same thing. Turning an absence found into an absence demonstrated is an academic speciality whose effectiveness stops at the first serious reviewer.

There remains what the four levels establish together, and which is more solid than any one of them taken in isolation. The question whether the use of a system modifies the capacity of the person who supervises it is today posed without shared criterion, without agreed stratification, without common observation horizon and without stable definition of what is being sought. One cannot settle it, not for want of data but for absence of instrument: a discipline endowed with quality indicators among the best standardised in all of medicine cannot manage to decide whether its assistance tool depreciates its practitioners. **The interesting problem is no longer whether deskilling exists. It is that we have neither common estimand, nor common protocol, nor common horizon allowing us to decide when it exists, in whom, at what speed and with what reversibility.**

The mechanism, for its part, has been theorised for forty-three years. Lisanne Bainbridge, in "Ironies of Automation" (Automatica, 19(6), 1983), established that automation removes routine operation from the operator, the very thing that maintains the competence required for the exceptional cases, which are what automation leaves to them. Cognitive delegation is nothing new: assisted diagnosis, air traffic control, expert systems and process control have been offloading cognitive tasks for decades. With generative AI, Bainbridge's paradox extends at scale to linguistic, analytical and normative tasks whose ground truth is often incomplete, costly or absent. A change of terrain rather than of nature, and the new terrain is the one where measurement is hardest.

Two recent works are consistent with this reading without demonstrating it. Lee and co-authors (Microsoft Research and Carnegie Mellon, CHI '25, doi:10.1145/3706598.3713778) surveyed 319 knowledge workers on 936 real-world uses: the higher the confidence in the system, the less critical thinking is deployed, the converse effect holding for self-confidence. The study is self-reported and measures perceptions, not performances. Kosmyna and co-authors (arXiv:2506.08872, non-peer-reviewed preprint, n = 54) report the lowest EEG connectivity in the assisted group and an inability of participants to quote their own output; a documented methodological critique contests its sample size, reproducibility and signal processing (Stanković et al., arXiv:2601.00856), and the first is not to be cited without the second.

Lin and co-authors formulated the general observation earlier and more sharply than I would have (arXiv:2602.00854, January 2026): assisted performance improves while the user's internal models degrade, which they name the Capability-Comprehension Gap, and they state the necessity of a minimum comprehension threshold below which supervision is no more than a procedure. They pose the question of the threshold without settling it empirically, and address neither the dynamics that lead to it, nor the organisational scale, nor the regulatory instrumentation. That is the space this text occupies.

3. The model

The naive reading of the Polish result consists in saying that AI makes people worse. That is true, banal and unusable. To make it an object of governance, one must stop treating oversight competence as a single block.

Four distinct capacities hide behind the word. Production: performing the task oneself. Verification: establishing whether an output is correct by confrontation with a reference. Recovery: taking back control when the system fails, under time constraint. Escalation: recognising that one does not know and knowing to whom to hand over. They depreciate neither at the same rate nor under the same conditions, and they are not mobilised by the same regimes:

$$H_{nominal} = f(C_V, , C_E, A\ T)\ H_{dégradé} = f(C_R, , C_P, , C_E\ A\ T)$$

where A is formal authority and T the time available to intervene. Production capacity does not condition routine oversight: one can contradict without knowing how to redo, through detection of internal inconsistencies, calibration of uncertainty, identification of out-of-distribution cases, recourse to a second independent system, or return to primary sources. Knowing how to redo and knowing how to contradict are not the same competence. But production capacity becomes critical again in degraded regime, when what is required is no longer to judge an output but to produce an acceptable substitute oneself. Escalation figures in both regimes, and for different reasons: recognising in normal operation that a case exceeds one's competence, recognising in degraded mode that one will not take back control alone. This is the classic distinction between nominal operation and degraded mode, and it lifts a tension that the previous version of this text left open: one can at once exclude production from routine oversight and want to measure it as an indicator of resilience.

In "Human oversight is not a presence", I argued that supervision is not a state but a throughput. Competence is only one factor in that throughput, and T is the one organisations compress first.

What governs degradation is not the nature of the tool but the availability of a reference external to the thing being verified. Classical automation offloaded the gesture while leaving the measuring instrument intact: the autopilot drifts, the altimeter says so. There is not, on one side, the tools equipped with an altimeter and, on the other, generative AI deprived of one. There is a gradient, and it follows the task, not the technology.

Task	External reference	Exposure to substitution
Numerical computation	strong	low
Testable code	strong	low to moderate
Document extraction	fairly strong	moderate
Clinical diagnosis	partial	moderate to high
Strategic analysis	weak	high
Normative reasoning, arbitration	weak	high

This table describes an exposure, not a risk, and its ordinality is illustrative, not measured. The nuance is decisive. The available reference is not the consulted reference. A highly verifiable task produces massive deskilling if no one ever consults the reference, and the developer who accepts a suggestion without running the test is the daily example of it. Conversely, a weakly verifiable task can remain healthy if the human works against the system rather than with it. The risk therefore depends on four terms and not one: the rate of delegation, the external verifiability of the task, the effective engagement in verification, and the quality of the feedback received.

The rate of delegation itself covers opposite regimes: substitution when the system does the work instead; augmentation when it extends the space of reasoning; tutoring when it produces feedback that increases competence, which does happen and must be said; complacency when engagement decreases without anything being learned. Two organisations using the same model in the same proportion can produce opposite trajectories. The determining variable is not the volume of use but the architecture of the interaction, which brings to light an object that no one regulates: the design of delegation.

Subject to that reservation, the dynamics can be written. Let C be the independent capacity of an operator, bounded by a ceiling C_{max}, let δ be the share of cases handled without autonomous engagement, λ the rate of depreciation through non-practice and μ the rate of acquisition through practice:

$$\frac{dC}{dt} = -\lambda\,\delta\,C \;+\; \mu\,(1-\delta)\,(C_{max} - C)$$

Competence becomes what it is: a bounded stock, maintained by exercise and eroded by non-practice. The model remains a toy with no predictive value, but it renders visible what regulatory reasoning assumes, namely that λ equals zero. Above all it brings out the question I was not posing: not whether competence declines, but at what speed it is restored relative to the speed at which it depreciates.

- If $\tau_{recovery} \ll \tau_{decay}$, the problem is one of preparation before intervention, serious but manageable.
- If $\tau_{recovery} \gg \tau_{decay}$, there is hysteresis, and only then is one entitled to speak of debt.

No longitudinal measurement of these two constants has been identified, neither in the systematic review cited above nor in the protocols of the two trials that oppose each other. That is the central empirical hole in this dossier, more important than the Polish figure, and confirming it requires the dedicated negative search announced in section 2.

There remains the fact that δ is not exogenous, and this is what completes the transformation of the arrangement into a dynamic system rather than a decay function. The rate of delegation itself depends on competence, on trust, on workload and on the perceived performance of the model. A less competent operator delegates more, which makes them less competent. An expert also delegates more, but for the inverse reason, because they are better at recognising what can be delegated. The same usage statistic therefore covers both a positive feedback loop and a rational allocation, and nothing in a dashboard distinguishes them.

To which is added a second loop, rarely formulated: in certain deployments, and presumably increasingly, human validations, overrides and corrections then serve to train, route, tune or evaluate the system. Elsewhere, this feedback is simply logged without feeding anything, and the loop does not close. It is therefore a class of architectures, not a universal property.

$$IA_t \to \delta_t \to C_{t+1} \to Supervision_{t+1} \to \acute{R}esultats_{t+1} \text{ avec } C_t \to \delta_t \text{ et } Supervision_t \to IA_{t+1}$$

Where it does close, the system progressively modifies the judge who then contributes to producing the data by which both the quality of the system and that of its supervision are evaluated. If the controller becomes complacent, the signal degrades with nothing to signal it. None of the observable quantities is then identifiable in isolation: an override rate is jointly produced by model quality, human capacity, trust, the interface and workload, and nothing allows it to be attributed to one rather than another. This is no

longer deskilling. It is a co-evolution of the controller and the controlled, in which measurement itself becomes endogenous.

4. The change of scale

The relevant unit of analysis is not the operator but the organisation, and collective competence is not the sum of individual competences: : $C_{org} \neq \sum_i C_i$.

An organisation can retain perfectly competent experts and lose its oversight capacity because they are too few, no longer in the loop, called upon too late, or present somewhere other than at the point of decision.

Above all, it depends on two stocks and not one: the experts in place, and the mechanism that produces the next ones. That mechanism runs through the tasks the organisation has just delegated. The senior remains competent for another five years. The junior never had to learn what the model now does in their place, and will therefore not become the senior they would have been. An organisation can thus satisfy its competence threshold today while having already lost the capacity to satisfy it tomorrow. The stock is compliant; the flow that replenishes it no longer is.

The signature of this phenomenon is that of invisible technical debt. Nominal performance remains constant, sometimes improves, and nothing appears in the indicators as long as the old component works. What decreases silently is resilience, that is, the capacity to absorb a failure. In reliability vocabulary: human competence becomes a non-redundant capability whose mean time to recovery increases while nominal operations remain excellent, and no one monitors the mean time to recovery of a skill. When the seniors leave, organisational competence does not decline steadily, it goes over a cliff. A linear depreciation at the individual level produces a discontinuity at the collective level, and it is the discontinuity that makes the accident.

It is here that the accounting metaphor must be held with rigour, because ordinary vocabulary conflates two objects. An asset depreciates; a debt accumulates. Competence is the stock, cognitive substitution is its depreciation mechanism, and the debt is neither one nor the other: it is a gap, between required capacity and available capacity.

$$D_t = \max\left(0,\ C_{requise,\,t} - C_{disponible,\,t}\right)$$

Written this way, it recalls that it possesses two sources and not one. It grows when available capacity falls, which is what section 3 describes. It also grows when required capacity increases, at rigorously unchanged competence, for example when a system is extended to more critical cases or to a wider perimeter. An organisation can take on debt without depreciating. The vocabulary then stabilises: the non-assisted performance test

becomes an impairment test, and recurrent practice a maintenance expense. There remains the question section 6 must settle, and the only one that counts: how much, exactly, is needed.

One last mechanism completes the picture and it is the most counter-intuitive. The better the system performs, the fewer reasons the human has to contradict it. The less they contradict it, the less they exercise their independent judgement. The less that judgement is exercised, the less available it is for the rare case on which the system fails. Yet it is exactly that rare case which motivates the supervision requirement. **As AI improves, human oversight becomes less often necessary and more difficult precisely at the moment when it becomes necessary.**

5. The regulatory gap

Article 14 of Regulation (EU) 2024/1689 requires high-risk systems to be designed so as to be effectively overseen by natural persons. Its paragraph 4 enumerates what those persons must be able to do: understand the capacities and limits of the system, correctly interpret its output, decide not to use it, disregard it, reverse it, interrupt the system. Point (b) enjoins remaining aware of the tendency to rely automatically on the output, designated in so many words as automation bias, the only place in the regulation where a precise cognitive phenomenon is named. Article 26(2) requires the deployer to entrust supervision to persons possessing the necessary competence, training and authority.

The arrangement is coherent and rests on an unstated assumption. Remaining aware of a bias is not retaining the capacity to correct it.

The gap must be formulated with precision, because its broad version would be false. To maintain that the regulation does not require the maintenance of competence would be imprudent: a contradictor would functionally reconstruct such an obligation from the deployer's duties, from the management system, from post-market surveillance or from sectoral standards, and they would have good arguments. **What is missing is not a norm, it is an instrument: the regulation requires competent supervision without defining any metric of recency for that competence, any minimum threshold of independent performance, or any periodic test establishing that the supervisory capacity remains recoverable.** The ISO/IEC 42001 standard does not fill the gap: its clause 7.2 reproduces the generic architecture of management systems — determine the required competences, train, evaluate the effectiveness of the training, retain the evidence. What the auditor examines there are qualification files and training records; nothing in the clause calls for a measurement of the operator's performance deprived of the tool. I do not conclude from this that no such measurement exists anywhere, which would require the negative search described above, but that the text serving as the international reference for an AI management system does not ask for it. Conformity attests to a title; it does not attest to a present fitness.

The discipline directly concerned by the Polish figure, for its part, did not wait. The European Society of Gastrointestinal Endoscopy published in 2026 a position arising from a formal Delphi consensus, defining a curriculum for the acquisition and maintenance of the competence necessary for the use of AI in endoscopy (Mori Y., Kopylov U., Sinonquel P. et al., Endoscopy 2026;58:202-210). Its recommendations show where the consensus is going: secure competence in standard endoscopy first, acquire the fundamentals of AI, recognise and mitigate the cognitive biases of human-machine interaction, avoid overconfidence in clinical decision-making, and continuously monitor quality indicators throughout the integration of the system. The sector produced in eighteen months what the horizontal layer does not contain. This is a counter-example to the broad version of my criticism and a confirmation of its narrow version: supervisory competence is indeed a regulable object, the demonstration has been made, and it has been made only where a professional community has taken hold of it discipline by discipline.

One sector, moreover, encountered the same problem forty years earlier and solved it in the reverse order. Aeronautics first held the licence to be a fitness, then observed that it was not. Section 61.57 of Title 14 of the Code of Federal Regulations prohibits a pilot from carrying passengers unless they have performed three take-offs and three landings within the preceding ninety days, as sole manipulator of the controls. This is not a requirement of a diploma, it is a requirement of recent and unassisted practice — enforceable, dated, verifiable. In January 2013, the Federal Aviation Administration published SAFO 13002, which states that continuous use of autoflight systems does not reinforce a pilot's knowledge and skills in manual flight and can degrade their ability to recover the aircraft rapidly from an undesired state. In May 2017, apparently judging that the message had not got through, the same administration republished the text under the title SAFO 17007, adding the word "proficiency".

Aviation therefore possesses a recency requirement, an explicit doctrine on degradation through automation, and a periodic arrangement that tests non-assisted performance. Gastrointestinal endoscopy possesses a competence-maintenance curriculum adopted by consensus. Horizontal AI governance possesses an obligation to remain aware.

What aeronautics eventually admitted is that control is not external to the system controlled. AI law reasons in terms of a static architecture of responsibilities: it assigns roles and verifies their existence at a given instant. It would need to reason in terms of a closed-loop dynamic system, in which oversight capacity is an endogenous state. This corrects an earlier position. Under the name Authority-Influence Gap, I had distinguished formal authority from effective cognitive control, a distinction others formulate as the gap between nominal supervision and effective supervision (arXiv:2604.00081). That gap is not an observed state, it is a trajectory: authority is stable by juridical construction, the capacity to exercise it follows the dynamics of section 3, and the gap opens mechanically. As for human oversight, it figures on the compliance balance sheet as a permanent asset,

non-amortisable, never subjected to an impairment test, and whose use is exactly what the deployment is intended to reduce.

Regulation (EU) 2026/1744, which entered into force on 27 July 2026, postponed to 2 December 2027 the application of the main obligations for stand-alone high-risk systems under Annex III, without modifying the substance of Article 14. Eighteen additional months of delegation before the asset is inventoried.

6. The instrumentation

What precedes does not establish the thesis, it establishes its plausibility and the absence of an instrument. Four limits must be held. The empirical base is thin and divergent: two serious trials conclude in opposite directions, including on experienced populations; a systematic review describes a heterogeneous and largely observational literature; and nothing today says which of the two results describes the general case. A result obtained on a perceptual act does not transport mechanically to analytical reasoning, and that is an extrapolation named as such. The constants of depreciation and restoration are unknown, the observed horizons short whereas the thesis bears on a cumulative phenomenon. Finally, the position defended is easier to state than to refute, which is the reproach addressed to the existential discourse in the opening: I do not claim to escape it, I claim to supply the means of getting out of it.

One must begin with the question that no doctrine of deskilling ever poses and which conditions all the rest: how much independent capacity must be preserved. Nothing obliges an organisation to preserve one hundred per cent of its historical autonomous capacity. It no longer knows how to etch its microprocessors, repair its lifts or compute a payroll by hand, and that is perfectly rational. Not every depreciation is a debt. The debt appears only at the crossing of a threshold:

$$C_{disponible} < C_{requise}\left(\acute{\text{r}}\text{esilience}\right)$$

And six parameters conceptually determine its level, none of which is cognitive: criticality of the task, frequency of system failures, detectability of those failures, time necessary for recovery, availability of independent alternatives, reversibility of the damage. What must be defined is therefore not a level of human competence, it is a minimum independent capacity for resilience: the capacity the organisation must retain in order to detect a failure, interrupt the system, maintain a safe state and restore an acceptable service. It is the exact analogue of the minimum viable service in IT resilience, transposed to competence.

This displacement lifts most of the objections addressed to doctrines of competence maintenance: it obliges the retention neither of all former competences nor, in each supervisor, of the capacity to redo everything, since competence can be distributed and escalation organised, and the threshold is proportionate to risk rather than uniform. The

objective is not to preserve maximum human autonomy, it is to preserve the minimum autonomy necessary for the mastery of risk. It is distinguished, finally, from the comprehension threshold proposed by Lin and co-authors, which is epistemic and bears on the contestability of an output: the resilience threshold is operational and bears on the capacity to hold the system in a safe state without it. Both are necessary and neither substitutes for the other.

Next comes what must be measured.

- **First renunciation: the single indicator.** The override rate seems the obvious candidate, and it is worth nothing on its own: a decline may signify an improvement of the model, a degradation of the controller, excessive trust, an increased workload, an organisational cost of contradiction, an interface that discourages disavowal, a selection of simpler cases, or time pressure. An indicator that admits eight competing explanations is not an indicator. What must be built is a panel, that is to say an observability of human oversight: blinded non-assisted performance, override rate, share of correct overrides, false-disavowal rate, calibration between stated confidence and accuracy, time to detection, capacity for independent justification, and performance after prolonged interruption of assistance, which is the only means of estimating τ_{recovery}.
- **Second point: split the protocol in two.** Measuring the operator without the system answers the question whether they still know how to do it, which is not the regulatory question. Whether they know how to supervise under real conditions is a different estimand. What is needed, therefore, is a test of autonomous competence, which measures the stock, and a test of supervisory effectiveness, which injects into a controlled environment a defined proportion of erroneous outputs and measures detection rate, correction rate and time to detection. A form of red teaming applied not to the model but to the supervisor.

Still, the right errors must be injected, failing which the test measures the trivial. A crude hallucination teaches nothing. A false answer, perfectly argued, correctly sourced except on one discreet premise, teaches a great deal. The protocol must therefore stratify:

$$E = \{ \text{crude, plausible, argued from authority, omission, framing, correlated} \}$$

And produce not a score but a probability of detection conditioned on the difficulty and the type of error. An omission and a factual untruth may present the same nominal difficulty and call upon different competences: what one obtains is therefore not a curve but a surface of supervisory performance, which is consistent with the refusal of the single indicator. One then obtains what compliance lacks today: not the attestation that a human was present, but the characterisation of what they detect and what escapes them. It is an experimental programme and not a contractual clause, and that is precisely what makes it useful.

The aeronautical transposition must not, however, be literal. Requiring a volume of cases handled without assistance would be premature: ten trivial cases are not worth one critical case, and the events one must know how to detect are often too rare to be encountered naturally. What aviation teaches is not a raw quota but a requirement of recurrent proficiency, complemented by simulation when natural exposure is insufficient. Duly noted: synthetic cases, rare scenarios and injected errors become the instrument for maintaining human competence in the face of AI. Simulated populations, long debated as a substitute for data, find here a less contestable function.

Which imposes a condition that all the preceding reasoning makes obligatory. If the supplier of the system also produces the test cases, the injected errors and the scoring scale, one reconstitutes exactly the coupling one claims to be measuring: the evaluation apparatus becomes an organ of the system evaluated. The independence of the test with respect to the system tested is a design property, not a good practice. It is obtained through distinct models or suppliers, externally validated case sets, an external ground truth, or synthetic generation followed by independent validation. Failing that, one asks the model to write the examination that will verify whether the human can spot its errors, which would constitute a rather creative evaluation protocol.

Finally, one must say what this costs, since that is the point on which a chief financial officer will decide. A minimum independent capacity is not free: cases handled without assistance, periodic simulations, double readings, adversarial exercises, expert time, back-up capacity. Its maintenance constitutes a recurrent resilience expense, and that expense belongs to the total cost of automation, not to the overheads of compliance.

It remains to arrange the objects in the right order. Extinction, dispossession and cognitive dependence are not three comparable risks: the first is a consequence, the second a state, the third a mechanism, and aligning them would amount to committing the error reproached in the opening. The correct form separates three levels. Mechanisms, among them cognitive depreciation, automation bias and contamination of the supervisory signal. An intermediate state, the loss of effective supervision. Consequences, from undetected errors through to catastrophe. Extinction occupies the last cell of the third column, with an undecidable probability. Cognitive substitution occupies the first cell of the first, with a mechanism documented since 1983, two trials that contradict each other for want of a common protocol, a systematic review that records the heterogeneity of the field, a regulatory precedent in a neighbouring sector, and a professional curriculum that proves the thing is regulable. Working on the first column is not renouncing the third, it is the only way of getting there.

Human oversight is not a guarantee that one enters into the file. It is a state variable of the system it is supposed to control, and the question is not whether it declines. The question is what level must be held, and having an instrument to establish that it is. **We shall keep the authority. It is not necessary for the machine to take control if we cease to maintain what would allow us to exercise it.**