

Humans don't leave the loop

They move up a level: where the accountable expert stands when agents do the work and the checking

Falk Uebernickel

University of St.Gallen (HSG)

Human thinking, AI-assisted writing: the ideas and the argument are the author's; the text was written together with AI.

Abstract. “Human in the loop” settles whether people oversee AI, but not on which level they do so. This note proposes four levels of oversight: the work, the output, its verification, and the rules that say what counts as right. Across three stages of AI maturity, which coexist in most organisations, the accountable expert stands on a different level. Yesterday, experts oversaw people and saw the work. Today, agents draft at volume and the expert signs off every output; in our observation, this is where most organisations using AI at volume stand, and where the bottleneck sits. Tomorrow, as a hypothesis, oversight agents check each output against explicit rule sets, and the expert moves up a level: writing the rules and checking the checkers. We set out what this move requires and where it stops, how the expert's job and its measure of success change, and five implications for leaders. The note continues Uebernickel (2026).

1 From whether to where

“Human in the loop” is reassuring and imprecise. A reviewer who glances at an output before it leaves the building is in the loop, and so is the expert who decides what an agent may do at all. The phrase covers both and says nothing about the difference, yet that difference decides whether oversight holds when volume grows.

In *Every task can be AI-supported* (Uebernickel, 2026) we argued that the useful question about AI at work is not *whether* a task can be supported but *how*, and that the form of support depends on how its output can be verified. If the form of support follows verifiability, then as verification becomes cheaper, the place where human oversight sits must move as well. The question becomes *where* the accountable expert stands: how far from the work, and responsible for what.

2 Four levels of oversight

We distinguish four levels at which a human can oversee work, ordered by their distance from the work itself:

- **Level 1 · Work.** *How was it done?* The overseer watches the process: sources, steps, judgement calls.
- **Level 2 · Output.** *What was delivered?* The overseer decides whether the result goes out.
- **Level 3 · Verification.** *Is the output correct?* The overseer checks the result against its source and against the criteria that apply. Quality and correctness sit here.
- **Level 4 · Rules.** *What counts as right in the first place?* The overseer sets the criteria. Two rule sets are worth separating: *must rules* say what is never allowed (law, compliance, contractual obligations); *should rules* say what leadership actually wants (price corridors, segment focus, positioning no-gos, escalation thresholds).

Each level takes its yardstick from the level above: verification without explicit rules checks

COUNTERPOINT

Experts may not be able to write what they know. Nisbett and Wilson (1977) reviewed evidence that people have little introspective access to the processes behind their judgements and report plausible theories instead. An expert's rule set may describe a theory of the work rather than the work. Which rules get written down: those that guide the judgement, or those that sound like it?

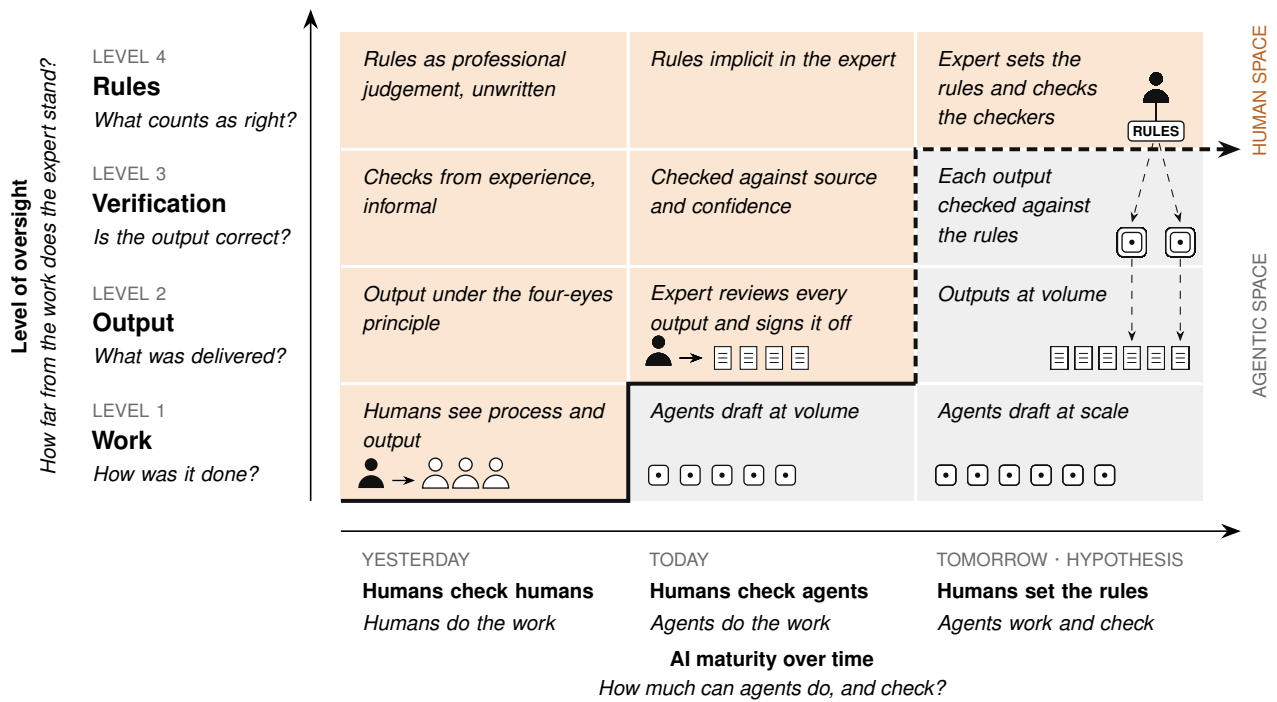


Figure 1. Four levels of oversight across three stages of AI maturity. The thick line marks where the accountable expert stands; below it, agents have taken over. Thin line from the expert to the rule set: setting the rules; dashed arrows: the oversight agents' checks. Dashed: hypothesis, which requires criteria set up front (frontloading). Filled figure: accountable expert; outlined: humans doing the work; squares: agents; double-framed: oversight agents. Own proposal; sign-off and frontloading after Hötter (2026).

against whatever the reviewer happens to have in mind. Level 4 says what counts as correct, Level 3 whether it holds in a given case (Figure 1).

3 Three stages: where the expert stands

The columns of Figure 1 are stages, not dates: they coexist, often within one firm.

Yesterday: humans check humans. Experts oversaw people: they saw the work and the output, applied the four-eyes principle and judged from experience. The rules were professional judgement, largely unwritten; the volume was human-sized, so this worked.

Today: humans check agents. Agents draft at volume, and the expert checks every output against its source and signs it off, combining Levels 2 and 3 in one act: the Sign-Off pattern described by Hötter (2026). In our observation, this is where most organisations using AI at volume stand, and where the bottleneck sits, because expert time does not scale with output. Two weaknesses follow. The rules sit in the expert's head, so they cannot be shared, versioned or challenged. And reviewers who mostly see correct items come to rely on them and miss the rare error (Parasuraman and Riley, 1997; Skitka et al., 1999); the sign-off drifts into a rubber stamp.

Tomorrow: humans set the rules (hypothesis). Oversight agents check each output against its source and against two explicit rule sets, must rules and should rules; in the design proposed here, each rule set has its own oversight agent. The expert moves to Level 4: writing the rules, and checking the checkers by sampling outputs with their own judgement, comparing the checkers' verdicts with known answers, and watching for drift. Written principles already guide AI critique of AI in model training (Bai et al., 2022). The tomorrow column is an organisational analogue of what alignment research calls scalable oversight (Bowman et al., 2022).

COUNTERPOINT

Expert judgement is a weak yardstick, at any level. Reviewing some hundred comparisons, Dawes et al. (1989) found that simple actuarial rules equalled or outperformed expert clinical judgement in nearly every study, and that letting experts override them tended to lower accuracy. When a checker and an expert disagree, who is more likely to be right?

4 What the move requires, and where it stops

Frontloading. The move presupposes frontloading: criteria fixed before anything is generated, a pattern Hötter (2026) observed in enterprise software sales. Where the evidence does not yet exist, as with a return-on-investment claim, there is no rule to check against; such tasks stay on the Proof Trail described in *Every task can be AI-supported*.

Reversibility. Where an effect cannot be undone, the gate before the effect stays human, whatever the checkers say. Autonomy should never run faster than the ability to intervene, as Carlos O. Fuentes put it in a comment on *Every task can be AI-supported*.

Independence of the checkers. An oversight agent from the same model family as the generator may share its blind spots. Studies of language models used as judges document position and verbosity biases and, with more limited evidence, self-enhancement bias (Zheng et al., 2023). Checkers should be simpler and more independent than the generators: rule-based where rules suffice, otherwise built on a different model or different data.

The rubber stamp moves up too. Every level inherits the failure mode of the one below it. An expert sampling the checkers' verdicts can drift into waving them through, so Level 4 needs the same measures as Level 2 (Section 5).

Where rule-writers learn. Writing a good rule requires knowing what goes wrong, and that is learned in the work being delegated. Bainbridge (1983) called this an irony of automation: the skills needed to supervise a system erode once the system does the work. Keeping experts fit for Level 4 means keeping a deliberate share of hands-on work.

The specification gap. Agents check against rules, not intentions; a rule can be met while its purpose is missed. Translating intent into rules, and noticing when the two diverge, becomes the expert's core task. In Tim Hourigan's distinction, governing behaviour and governing purpose are different jobs; only the first moves to the machine.

5 The expert's job changes

The three stages describe a role ladder. Yesterday the expert oversaw the people doing the work and judged the result, and success was the delivered output. Today the expert approves, and success should be the errors found, not the throughput of approvals. Tomorrow the expert sets the rules and checks the checkers, and success is what the rules catch and what still slips through. Moving up a level means fewer experts per output, not necessarily fewer experts: the work becomes writing rules, calibrating checkers and deciding exceptions.

In practice, roles are more often hollowed out than redesigned. The work moves to approving while the job profile, the incentives and the career path stay with producing. Three measures tell oversight from rubber-stamping at any level: the rejection rate (a rate near zero is a warning sign), the detection of deliberately planted errors, and how often a blind second judgement on a sample disagrees with the first. They are diagnoses, not targets; a measure that becomes a target ceases to be a good measure (Goodhart's law in the formulation of Strathern, 1997). The design principle behind all three: make saying no cheaper than saying yes.

6 Five implications for leaders

1. **Ask on which level, not whether.** Name the level the expert stands on, and the one the organisation wants.
2. **Write the rules down before adding reviewers.** Separate must rules from should rules. Version them like code, on the cadence of law and strategy rather than of model releases, and let people challenge them.
3. **Redesign the expert role, do not hollow it out.** Move job profile, incentives and

COUNTERPOINT

Fixing criteria before generation is exploitation by design. March (1991) shows that refining what is already known pays in the short run but crowds out the exploration that reveals what else is worth wanting. Oversight agents will enforce today's strategy faithfully. Where in this design does the organisation learn that its rules are wrong?

COUNTERPOINT

Different does not mean independent. Knight and Leveson (1986) had 27 programs written independently from one specification and found that they failed together far more often than independence would predict. A checker on another model may share the generator's blind spots because it shares the task. Is independence something you can build, or only measure?

COUNTERPOINT

A share of hands-on work may not be enough. In robotic surgery, Beane (2019) found that once the technology shrank trainees' hands-on role, approved training failed and only a minority built skill, through norm-bending shadow learning. A budgeted share of practice is still a diminished apprenticeship. Who will have seen enough failures to write tomorrow's rules?

COUNTERPOINT

Sampling may calibrate trust the wrong way. Dietvorst et al. (2015) found that people lose confidence in an algorithm faster than in a human after seeing both make the same mistake, and then prefer the inferior human. An expert who samples the checkers will see errors; the result may be abandonment rather than drift. Does checking the checkers build trust, or burn it?

time budget with the work.

4. **Check the checkers.** Sample, plant errors, track rejections. A sign-off that never says no is not a control, at any level. Where the EU AI Act applies to high-risk systems, its Article 14 requires effective human oversight (European Union, 2024); placing the human on Level 4 must be shown to be effective, not assumed.
5. **Keep a path for building expertise.** Tomorrow's rule-writers are trained in today's work.

7 Limitations

The tomorrow column is a hypothesis; the four levels, the two rule sets and the role ladder are a proposal, not a finding. Sign-off and frontloading were derived from enterprise software sales (Hötter, 2026) and not yet validated elsewhere. How much of a strategy can be written as checkable should rules is an open empirical question; the independence of oversight agents and the detection of drift without labelled data are open research problems. Not every task will move.

Acknowledgement. We thank Andrew J. Turner, Tim Hourigan, Marilyne Labasque (the term “expert in the loop”), Carlos O. Fuentes, Laurent Herbillon and the other commenters on LinkedIn whose objections shaped Section 4.

References

- Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint 2212.08073.
- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6):775–779.
- Beane, M. (2019). Shadow learning: Building robotic surgical skill when approved means fail. *Administrative Science Quarterly*, 64(1):87–123.
- Bowman, S. R., Hyun, J., Perez, E., et al. (2022). Measuring progress on scalable oversight for large language models. arXiv preprint 2211.03540.
- Dawes, R. M., Faust, D., and Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674.
- Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.
- European Union (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 12 July 2024.
- Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45:105681.
- Hötter, J. (2026). Verification asymmetry in enterprise SaaS sales: Human–AI interface design patterns for AI-based workflows. Master's thesis, Hasso Plattner Institute, University of Potsdam. Supervised by F. Uebernickel and T. Wuttke.
- Knight, J. C. and Leveson, N. G. (1986). An experimental evaluation of the assumption of independence in multiversion programming. *IEEE Transactions on Software Engineering*, SE-12(1):96–109.
- March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87.
- Nisbett, R. E. and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259.
- Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253.
- Skitka, L. J., Mosier, K. L., and Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006.
- Strathern, M. (1997). ‘Improving Ratings’: Audit in the British university system. *European Review*, 5(3):305–321.
- Uebernickel, F. (2026). Every task can be AI-supported. Working note, general edition, October 2026, <https://uebernickel.com/writing/every-task-can-be-ai-supported>.
- Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS 2023, Datasets and Benchmarks Track*.

COUNTERPOINT

A human on Level 4 may be a shield rather than a control. Surveying 41 policies requiring human oversight of government algorithms, Green (2022) finds that evidence suggests people cannot perform the oversight they demand, and that such requirements legitimise flawed systems without fixing them. Who answers when the rules were right and the checkers were wrong?