

Every task can be AI-supported

The question is how humans stay in charge, and how much oversight it takes

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. Debates about AI at work often ask *whether* a task should be touched by AI at all. This note takes a different position: there is hardly any area of knowledge work or operations left that AI cannot support. What differs between tasks is *how* AI supports them, and how much human oversight that form of support contains. Building on the asymmetry between generating and verifying an output (Cobbe et al., 2021; Wei, 2025) and on design patterns from recent thesis work (Hötter, 2026), we describe a matrix of generation effort and verification effort. Each quadrant is a form of AI support with its own built-in oversight: Autopilot, Sign-Off, Proof Trail and Co-Creation. What remains outside the matrix, Human only, is not a technical limit but a conscious choice. We illustrate the matrix with typical tasks from sales and operations, derive five implications for leaders, and suggest where to start. The examples are illustrative.

1 From whether to how

Many AI frameworks start with risk: the higher the risk, the more human involvement. Leaders draw their lines in a similar way. Strategy, people decisions, client relationships and safety stay with them; reporting, scheduling and routine decisions could be handed over. Lines like this are usually drawn along the question *can AI do this, or not?*

Our position is that this question has largely been answered. Almost any task can now be supported by AI, from compiling a supplier dossier to preparing a pricing decision. The useful question is *how* AI should support a given task, and how humans stay in charge while it does. A risk class does not answer this: it says how much human involvement a task needs, not what kind. Consider two tasks from business-to-business sales with similar stakes. A wrong answer in a security questionnaire can cost the deal. Yet the questionnaire is laborious to compile but quick to check, because every answer points to a policy, a certificate or a contract clause. A return-on-investment claim takes minutes to write, yet whether it holds becomes visible only months after go-live. AI can support both. It should support them in very different ways.

2 Two efforts, four ways AI supports the work

The matrix rests on one observation: for many tasks, checking an answer is far cheaper than producing it. Machine-learning research exploits this asymmetry when a separate verifier ranks generated solutions (Cobbe et al., 2021). Wei (2025) states it as a general rule, the verifier's law: the ease of training AI to solve a task grows with how easily the task can be verified. For organisations, the asymmetry has a second meaning. If AI makes generation cheap, the form of support is decided by how the result can be checked. Following Wei (2025), the matrix places two efforts on two axes (Figure 1):

- **Generation effort** (horizontal): how hard is it to produce the output? For example, the expertise required, the number of sources, the amount of context.
- **Verification effort** (vertical): how hard is it to check the output? For example, the time to check, the number of roles involved and the cost of correcting an error.

COUNTERPOINT

Risk is not a property of single outputs. Perrow (1984) located the gravest failures in systems that are interactively complex and tightly coupled, where small failures, each harmless on its own, combine in unforeseen ways. A verification axis scores each task on its own. Can a per-task matrix see the risk that arises between tasks?

COUNTERPOINT

Support may be thinner than it looks. Autor (2015) argues that the tasks hardest to automate demand flexibility, judgement and common sense, skills we understand only tacitly: Polanyi's paradox, we know more than we can tell. Where the hard part of a task cannot be articulated, AI supports the paperwork around it. Is every task supportable, or only its describable part?

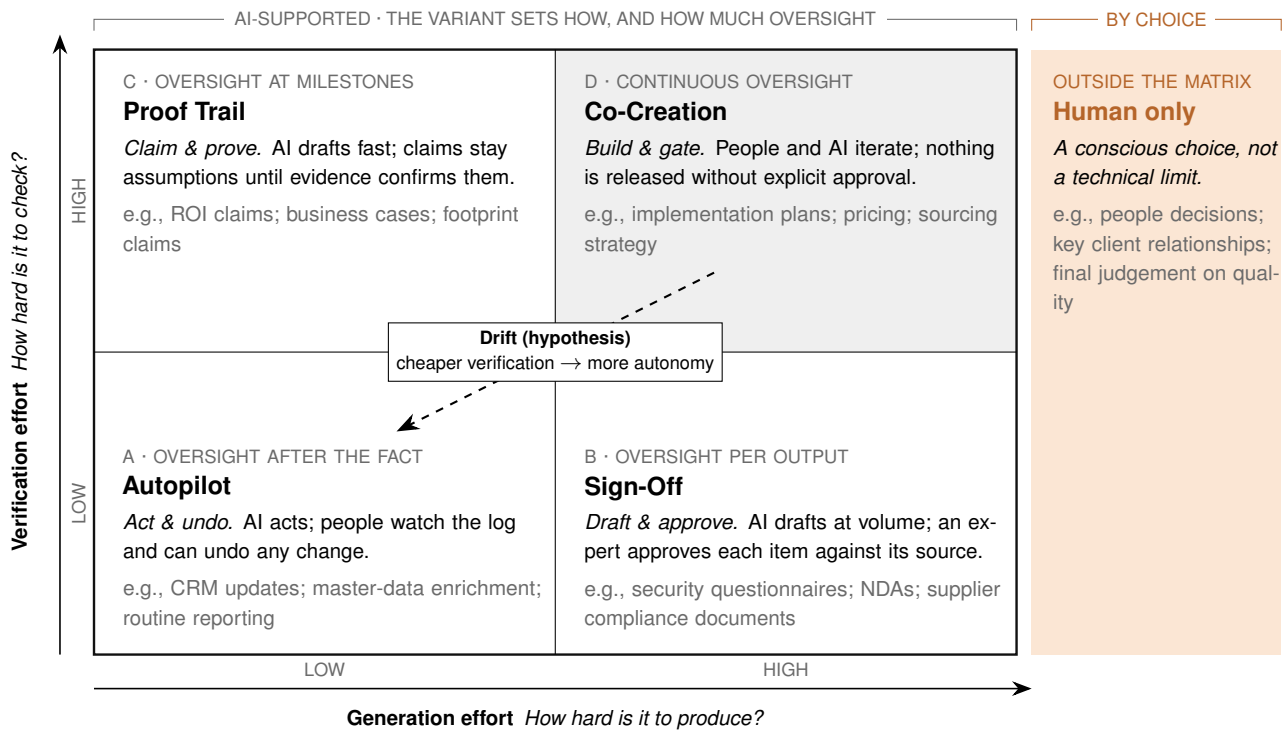


Figure 1. Four ways AI supports the work, with illustrative tasks from sales and operations. Every cell is AI-supported; what differs is the form of support and the oversight built into it. Placement depends on how a task is set up in a given organisation and is meant as a starting point for discussion. Adapted from Hötter (2026); axes after Wei (2025).

In Hötter (2026), each axis is scored with five binary criteria (0–2 low, 3–5 high). Risk is not discarded. It enters through the cost of correction, as part of verification effort. An action that cannot be undone is, by definition, expensive to verify after the fact.

A · Autopilot (low generation, low verification). AI acts; people supervise the log and keep the power to undo. This suits high-volume, reversible work such as updating CRM records, enriching master data or refreshing a delivery status. It fails when the undo button is missing. A customer-facing answer, once given, cannot be reversed: in *Moffatt v. Air Canada* (2024 BCCRT 149), an airline was held liable for what its website chatbot told a customer. Irreversibility moves a task up the verification axis, whatever its generation effort.

B · Sign-Off (high generation, low verification). AI compiles at volume; an expert approves each item against its source. Typical cases are the documentation trails that business already runs on: security questionnaires and non-disclosure agreements in sales, compliance evidence from suppliers, regulatory paperwork. Here AI removes most of the drafting work, and the risk shifts to the reviewer. People who approve mostly correct items drift into approving everything (Parasuraman and Riley, 1997; Skitka et al., 1999), and rare errors are missed far more often than common ones (Wolfe et al., 2005). A sign-off that never says no is not a control.

C · Proof Trail (low generation, high verification). Claims are cheap to write and slow to confirm: the return of a new tool, a forecast, a footprint reduction. AI can draft them in minutes. The oversight lies in what happens next: each claim stays labelled as an assumption, gets an owner and a date by which evidence is due, and is signed off by the accountable role once the evidence arrives. The failure mode is quiet. Assumptions harden into facts because nobody revisits them.

D · Co-Creation (high generation, high verification). Even strategic and design work, such as a pricing model, an implementation plan or a sourcing strategy under volatile prices, can be AI-supported. Here, people and AI build the result together, and explicit gates decide what leaves the room. Field evidence shows why the gates matter. In an experiment with

COUNTERPOINT

Checking against the source is quick only if the error is visible. In Kabir et al. (2024), 52% of ChatGPT answers to programming questions contained incorrect information; participants still preferred them 35% of the time and overlooked the misinformation 39% of the time. How quick is the check when fluency hides the fault?

COUNTERPOINT

The owner of a claim is rarely its best auditor. Across 258 transport infrastructure projects, Flyvbjerg et al. (2002) found cost underestimation better explained by strategic misrepresentation than by error. Who holds the evidence when the claimant holds the date?

consultants, AI raised quality on tasks inside its capability frontier but lowered correctness on a task just outside it, where consultants trusted it too readily (Dell'Acqua et al., 2023).

Human only: a conscious choice. If almost every task can be AI-supported, then the work that stays purely human is no longer a remainder that technology has not yet reached. It is a decision. An organisation may decide that decisions about people, its most important client relationships and the final judgement on quality remain human because the human act itself carries the value: craft, accountability, trust. Such a choice should be named, justified and revisited from time to time, rather than left as an unspoken hesitation.

3 Drift: human-in-the-loop as a transition

The matrix is not static. As verification becomes cheaper, through linked sources, structured outputs, automated tests or rule-based checks, tasks move down and to the left, towards Autopilot. In software engineering, where every change can be checked against an automated test suite, models resolve a growing share of real issues (Jimenez et al., 2024). We therefore treat per-item approval as a transition technology, as a hypothesis rather than a finding. Its consequence is that the human role changes rather than disappears: it moves from checking each output to defining what counts as correct and checking the checkers, the agenda known as scalable oversight (Bowman et al., 2022).

4 Five implications for leaders

1. **Ask how, not whether.** Start from the assumption that every process can be AI-supported. Before shelving a use case as “too risky”, ask how its output could be checked; the answer selects the variant.
2. **Make Human only a deliberate decision.** Agree, per function, which acts remain human, and write down why. An explicit, reviewed choice protects craft and client relationships better than a general hesitation.
3. **Budget expert time as the bottleneck.** Generation is becoming cheap; verification is not. Every Sign-Off use case consumes the hours of the people whose judgement the organisation depends on.
4. **Measure what the oversight catches.** Track what reviewers change and why, and occasionally plant known errors to see whether they are caught. An approval rate close to 100% is a warning sign, not a success.
5. **Write the rules down.** Separate what is never allowed (law, compliance, industry standards) from what leadership wants (company standards, service style). Written, versioned rules are the precondition for scaling oversight. Where the EU AI Act applies to high-risk systems, its Article 14 requires effective human oversight (European Union, 2024).

5 Where to start

We suggest building a portfolio of AI initiatives systematically rather than case by case. A first step can be a small strategy project that selects where to start, places every candidate use case on the matrix, and names the Human only acts for each function. Within that portfolio, the priority should be *augmentation*: use cases in which AI strengthens people's work while they keep the judgement and the accountability, as in Sign-Off, Proof Trail and Co-Creation. Starting there builds the review routines, written rules and measures of what oversight catches (Section 4) that any later move towards Autopilot depends on.

COUNTERPOINT

Co-creation may not stay co-creation. Noy and Zhang (2023) found that ChatGPT raised professionals' writing productivity substantially and, in the authors' reading, mostly substituted for worker effort rather than complementing skills; in the working paper, most participants who used it submitted its first output unedited. Is building together a stable form, or a slide into delegation?

COUNTERPOINT

Cheaper checks verify what they were written to verify. In a user study, Perry et al. (2023) found that participants with an AI assistant wrote significantly less secure code than those without, and were more likely to believe their code was secure. Code can pass its tests and still be unsafe. Does cheaper verification make autonomy safer, or only make drift feel safe?

COUNTERPOINT

Human only may be the customer's choice, not the firm's. Longoni et al. (2019) found that consumers were less likely to use healthcare from an AI provider, out of uniqueness neglect, the concern that AI cannot account for their own circumstances; resistance eased when AI supported, rather than replaced, a human provider. Who decides what stays human?

COUNTERPOINT

Augmentation carries its own loss. Across 106 experiments, Vaccaro et al. (2024) found that human-AI combinations performed worse on average than the best of humans or AI alone: gains when humans outperformed the AI, losses when the AI outperformed humans. Which Sign-Off tasks would the AI do better alone?

6 Limitations

That every task can be AI-supported is a position, not an empirical finding, and support does not mean automation. The matrix was derived from enterprise software sales (Hötter, 2026) and has not yet been validated in other domains such as manufacturing or services. The task examples are illustrative and do not describe any particular organisation. A portfolio as proposed above would also test the matrix: placing, running and measuring a few use cases over two to three quarters.

References

- Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, 29(3):3–30.
- Bowman, S. R., Hyun, J., Perez, E., et al. (2022). Measuring progress on scalable oversight for large language models. arXiv preprint 2211.03540.
- Cobbe, K., Kosaraju, V., Bavarian, M., et al. (2021). Training verifiers to solve math word problems. arXiv preprint 2110.14168.
- Dell’Acqua, F., McFowland III, E., Mollick, E. R., et al. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working Paper 24-013, Harvard Business School.
- European Union (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, OJ L, 2024/1689, 12 July 2024.
- Flyvbjerg, B., Holm, M. S., and Buhl, S. (2002). Underestimating costs in public works projects: Error or lie? *Journal of the American Planning Association*, 68(3):279–295.
- 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.
- Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. (2024). SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*.
- Kabir, S., Udo-Imeh, D. N., Kou, B., and Zhang, T. (2024). Is Stack Overflow obsolete? An empirical study of the characteristics of ChatGPT answers to Stack Overflow questions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI ’24)*, page Article 935, Honolulu. ACM.
- Longoni, C., Bonezzi, A., and Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4):629–650.
- Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.
- Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253.
- Perrow, C. (1984). *Normal Accidents: Living with High-Risk Technologies*. Basic Books, New York.
- Perry, N., Srivastava, M., Kumar, D., and Boneh, D. (2023). Do users write more insecure code with AI assistants? In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS ’23)*, pages 2785–2799, Copenhagen. ACM.
- 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.
- Vaccaro, M., Almaatouq, A., and Malone, T. W. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303.
- Wei, J. (2025). Asymmetry of verification and verifier’s law. Blog post, <https://www.jasonwei.net>.
- Wolfe, J. M., Horowitz, T. S., and Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature*, 435(7041):439–440.