

# Healthy AI First, Episode 8: Measure Honestly, Try It Yourself (Series Finale)

## Sources and evidence for this episode

Miriam's department head wants a number by the end of the month: how much faster is the team with AI? Instead of one big shiny number, the episode follows Miriam, Daniel and Jonas for four weeks like a diary: week one just watching, in weeks two and three one works with the new rule and one with the old, they swap after a week, and week four is for evaluation. They measure with a dashboard of three gauges (minutes per change, energy at night, the comprehension conversation). Along the way: Goodhart's law, co-determination and data protection, a broken Thursday that an AI summary hides in the median, and one page for the budget meeting. At the end, your number one last time.

All links were checked on 17 October 2026.

## What we mention in this episode

---

### The First Temptation

- **Claim:** When a measure becomes a target, it ceases to be a good measure; the idea goes back to the economist Charles Goodhart, the well-known short form comes from the anthropologist Marilyn Strathern.

**Source:** Goodhart, C. (1975): Problems of Monetary Management: The U.K. Experience. Papers in Monetary Economics, Reserve Bank of Australia; Strathern, M. (1997): "Improving ratings": audit in the British University system. European Review 5(3), 305-321.

**Link:** <https://doi.org/10.1017/S1062798700002660> (overview with Goodhart's original wording: [http://en.wikipedia.org/wiki/Goodhart%27s\\_law](http://en.wikipedia.org/wiki/Goodhart%27s_law))

**Strength and limits:** A rule of experience from economics and anthropology, not an experiment. The example "rewarding merged changes" applies the principle.

### The Dashboard

- **Claim:** Minutes per change count only human time (clarifying, explaining, reviewing, fixing); "done" means criteria met, checks passed, and a human takes responsibility.
- **Source:** Basis of the series, chapter 14 ("human attention minutes per accepted change").
- **Strength and limits:** Explicitly a working hypothesis, not an established standard.
- **Claim:** Several gauges instead of one number: a few metrics from several dimensions, deliberately in tension, with at least one perception measure, reported only at team level; do not turn the metrics into targets or make teams compete.

**Source:** Forsgren, N., Storey, M.-A., Maddila, C., Zimmermann, T., Houck, B. & Butler, J. (2021): The SPACE of Developer Productivity. ACM Queue 19(1), 20-48; DORA: Software delivery performance metrics.

**Link:** <https://doi.org/10.1145/3454122.3454124> (<https://queue.acm.org/detail.cfm?id=3454124>) · <http://dora.dev/guides/dora-metrics/>

**Strength and limits:** Research-based measurement frameworks, not a single causal study. Not named in the audio; details only in the show notes.

## Fuel Gauge and Warning Light

- **Claim:** Energy at night on a scale of one to five is a local signal, not a diagnosis; NASA-TLX measures subjective workload across six dimensions; it complements the evening-energy scale but measures a different construct.  
**Source:** Basis of the series, chapter 14; NASA Human Systems Integration Division: NASA Task Load Index (TLX); Hart, S. G. (2006): NASA-Task Load Index (NASA-TLX); 20 Years Later. Proceedings of the HFES Annual Meeting 50(9), 904-908.  
**Link:** <https://www.nasa.gov/human-systems-integration-division/nasa-task-load-index-tlx/> · <https://doi.org/10.1177/154193120605000909>  
**Strength and limits:** The one-to-five scale is a convention of the series; NASA-TLX is an established workload instrument, not a fatigue diagnosis.
- **Claim:** The comprehension conversation checks a randomly drawn change: what must never happen, and what happens with a failure no test covers; self-confidence alone is not enough.  
**Source:** Basis of the series, chapters 14 and 15.  
**Strength and limits:** Proposal, not a validated method. A technically correct explanation is evidence of understanding; the conversation is a proposed check, not a complete proof.

## Week One: Just Watching

- **Claim:** Times and energy ratings per person are personal data; where there is a works council, a technical system that can monitor the performance or behavior of individuals requires co-determination in Germany.  
**Source:** Section 87(1) no. 6 of the German Works Constitution Act (Betriebsverfassungsgesetz); Federal Labour Court, decision of 9 Sep 1975, 1 ABR 20/74 (objective suitability for monitoring is enough).  
**Link:** [https://www.gesetze-im-internet.de/betrvg/\\_\\_\\_87.html](https://www.gesetze-im-internet.de/betrvg/___87.html) · <https://www.haufe.de/id/entscheidungen/bag-beschluss-vom-09091975-1-abr-2074-HI436774.html>  
**Strength and limits:** Statute and case law (German sources); rules differ by country. Not legal advice.
- **Claim:** Claude Code can, on request, record how long a human is actively using its interface, separating the seconds in which someone types and reads from those in which the agent works; a timestamp is still not attention.  
**Source:** Anthropic: Claude Code documentation, Monitoring.  
**Link:** <https://code.claude.com/docs/en/monitoring-usage>  
**Strength and limits:** Vendor documentation; measures time spent within the Claude Code interface, not attention or load. Git provides commit timestamps; reviews are recorded by the pull-request platform; together with calendars, these records do not yet give active human minutes. Work outside them must be added, overlaps must not be double counted, and remaining measurement gaps should be disclosed; errors can go in either direction.

## Switching Sides

- **Claim:** Week one baseline, weeks two and three switching conditions (crossover), week four evaluation; the order is balanced so that one easy or hard week does not decide the result. It is not perfectly clean, because learning effects carry over the swap.  
**Source:** Basis of the series, chapter 15.

**Strength and limits:** A common experimental design; in a small team a learning signal, not statistical proof. The football image is ours.

- **Claim:** The new rule comes from episode 4: one complex task at a time, at most one quiet background job, a fixed limit for changes waiting for review; success criteria and stop limits are set beforehand.

**Source:** Basis of the series, chapters 7 and 15; DORA: Work in process limits.

**Link:** <https://dora.dev/capabilities/wip-limits/>

**Strength and limits:** Pilot proposal; there is no universal percentage for success.

## Week Three

- **Claim:** Turning every saving straight into more work means measuring something else; with the same actual throughput and stable averages, more work in the system means a longer average lead time.

**Source:** Basis of the series, chapter 13; Little, J. D. C. (1961): A Proof for the Queuing Formula:  $L = \lambda W$ . Operations Research 9(3), 383-387 (covered in episode 4).

**Link:** <https://doi.org/10.1287/opre.9.3.383>

**Strength and limits:** A typical failure pattern according to the basis of the series; Little's law holds for stable systems on average.

## Week Four: The Smooth Summary

- **Claim:** A one-day outlier among a few dozen values barely shows in the median. The mean rises substantially because it includes the real repair cost; without a measurement error, that is not distortion. So also show the distribution and total human effort. An AI summary is not a review.

**Source:** Basic statistics; basis of the series, chapters 13 to 15 ("distributions and outliers instead of just an average").

**Strength and limits:** General statistics; the broken Thursday is invented.

## One Page for the Budget Meeting

- **Claim:** What was measured (recorded time, additional activities and known measurement gaps; energy and understanding from the people themselves), what showed up, what remains open, and which next decision is due.

**Source:** Basis of the series, chapters 15 and 16 (templates).

**Strength and limits:** A template; the result of Miriam's trial is invented.

## What Nobody Knows Yet

- **Claim:** It has long been known that vigilance is demanding; how good a review still is late in the afternoon after a day full of agent work we are not aware of robust direct measurements of it. It is also open whether AI frees up resources or the expected workload simply grows with it.

**Source:** Warm, J. S., Parasuraman, R. & Matthews, G. (2008): Vigilance Requires Hard Mental Work and Is Stressful. Human Factors 50(3), 433-441; basis of the series, chapter 17.

**Link:** <https://doi.org/10.1518/001872008X312152>

**Strength and limits:** Vigilance research as a basis; applying it to code review is an open research question.

## Your Number

- **Claim:** In episode 1, Daniel stood for sixteen experienced developers who estimated they were almost a quarter faster and were measured slower.  
**Source:** Becker, J., Rush, N., Barnes, E. & Rein, D. (METR, 2025): Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. Preprint.  
**Link:** <https://arxiv.org/abs/2507.09089> (summary: <https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/>)  
**Strength and limits:** Randomized study, 16 people, tools from early 2025.

## Tool tip: collect times automatically

---

- Claude Code monitoring via OpenTelemetry: enable with `CLAUDE_CODE_ENABLE_TELEMETRY=1` plus exporter configuration (off by default). The metric `claude_code.active_time.total` measures active time in seconds, attribute `type = user` (keyboard input) or `cli` (tool execution and AI responses). <https://code.claude.com/docs/en/monitoring-usage>
- Times from pull requests and commits (see episode 2): Git provides commit timestamps; reviews and merges are recorded by the pull-request platform. These are lead times, not active human minutes. [https://cli.github.com/manual/gh\\_pr\\_list](https://cli.github.com/manual/gh_pr_list) · <https://git-scm.com/docs/git-log>

## Further reading

---

- DORA: A history of DORA's software delivery metrics. Deployment rework rate was added in 2024 as the fifth metric (unplanned deployments resulting from a production incident, not the same as internal rework). <https://dora.dev/insights/dora-metrics-history/>
- Noda, A., Storey, M.-A., Forsgren, N. & Greiler, M. (2023): DevEx: What Actually Drives Productivity. ACM Queue 21(2), 35-53. <https://doi.org/10.1145/3595878>
- Levitt, S. D. & List, J. A. (2011): Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments. American Economic Journal: Applied Economics 3(1), 224-238. The famous factory story is shakier than its reputation. <https://doi.org/10.1257/app.3.1.224> (working paper: <https://www.nber.org/papers/w15016>)
- Paradis, E. et al. (2024): How much does AI impact development speed? An enterprise-based randomized controlled trial. 96 Google software engineers; best estimate about 21% less time, wide confidence interval. Preprint. <https://arxiv.org/abs/2410.12944>
- Basis of the series: Healthy AI First Software Development, version 2 (1 October 2026), chapters 12 to 17.

## Note

---

Healthy AI First is AI-researched and AI-generated, narrated in Björn's voice, recreated with AI, and reviewed by him.

Miriam, Daniel, Jonas and the department head are fictional characters used as illustrations. In episode 1, Daniel stood in for the participants of the METR study; from episode 2 on he is an invented colleague

and did not take part in the study. The course and results of Miriam's trial (the hard week two, the broken Thursday, the comprehension conversation) are invented as well.