Healthy AI First, Folge 1: Gefühlt schneller

Quellen und Belege zur Folge

Sechzehn erfahrene Entwickler erwarten, dass KI sie fast ein Viertel schneller macht. Gemessen werden sie langsamer, und hinterher halten sie sich trotzdem für schneller. Die Auftaktfolge erzählt, warum unser Gefühl bei der Arbeit mit KI ein schlechtes Messinstrument ist, was andere Studien dagegenhalten (große Feldexperimente, Branchenberichte, Befragungen), was die Arbeit mit KI kosten kann (mehr Prüfen, Ermüdung, weniger Lernen) und wohin die Serie in den nächsten sieben Folgen geht.

Alle Links wurden am 17.10.2026 geprüft.

Was wir in dieser Folge erwähnen

Das Experiment, Die erste Wendung, Die zweite Wendung

- **Aussage:** Erfahrene Open-Source-Entwickler erwarteten mit KI etwa 24 % Zeitersparnis, brauchten gemessen etwa 19 % länger und schätzten hinterher, sie hätten etwa 20 % Zeit gespart.
  Quelle: 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 (Zusammenfassung: https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/)
  Einordnung: Randomisierte Studie mit echten Aufgaben (16 Personen, 246 Aufgaben), daher starke Evidenz für genau diese Situation; aber kleine Gruppe, erfahrene Leute im eigenen Projekt und Werkzeuge von Anfang 2025.

Das Experiment zerbricht

- **Aussage:** Die Folgestudie mit 57 Entwicklern und über 800 Aufgaben lieferte keine verlässliche Zahl, weil viele nicht mehr ohne KI arbeiten wollten und 30 bis 50 % der Entwickler angaben, manche Aufgaben deshalb gar nicht eingereicht zu haben.
  Quelle: Becker, J., Rush, N., Cunningham, T., Rein, D. & Mahamud, K. (METR, 24.02.2026): We are Changing our Developer Productivity Experiment Design.
  Link: https://metr.org/blog/2026-02-24-uplift-update/
  Einordnung: Die Autoren nennen ihre eigene Schätzung ausdrücklich ein unzuverlässiges Signal (Selektion, schwierige Zeitmessung bei parallelen Agenten). Frühe Evidenz, Blogbericht.

Die Gegenstimmen

- **Aussage:** In drei randomisierten Feldexperimenten bei Microsoft, Accenture und einem weiteren Großunternehmen schlossen Entwickler, die den KI-Assistenten tatsächlich nutzten, etwa ein Viertel mehr Aufgaben pro Woche ab (+26,08 %); weniger erfahrene profitierten stärker.
  Quelle: Cui, K. Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S. & Salz, T. (2026): The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers. Management Science.
  Link: https://doi.org/10.1287/mnsc.2025.00535 (Working Paper: https://doi.org/10.2139/ssrn.4945566)

Einordnung: Große randomisierte Feldstudie (4.867 Personen), starke Evidenz; gemessen werden abgeschlossene Pull Requests, nicht Qualität. Der Effekt der Nutzung wird mit zusätzlichen statistischen Annahmen geschätzt, die einzelnen Experimente streuen stark.

- **Aussage:** Die DORA-Berichte von Google beschreiben KI als Verstärker: Wo mehr KI genutzt wird, wird mehr ausgeliefert, die Stabilität leidet aber; gut organisierte Teams gewinnen, schwache Strukturen werden schneller sichtbar.
  Quelle: DORA und Google Cloud (2025): State of AI-assisted Software Development; DORA und Google Cloud (2024): Accelerate State of DevOps Report.
  Link: https://research.google/pubs/dora-2025-state-of-ai-assisted-software-development-report/ · https://dora.dev/insights/dora-2025-year-in-review/ · https://services.google.com/fh/files/misc/2024_final_dora_report.pdf
  Einordnung: Große Befragungen (2025 fast 5.000 Personen); zeigen Zusammenhänge, keine Ursachen. Selbstauskunft und Selbstselektion.
- **Aussage:** Im Frühjahr 2026 schätzten mehrere hundert technische Fachkräfte den Wert ihrer Arbeit mit KI typischerweise auf das 1,4- bis 2-Fache.
  Quelle: Becker, J. (METR, 11.05.2026): Measuring the Self-Reported Impact of Early-2026 AI on Technical Worker Productivity.
  Link: https://metr.org/blog/2026-05-11-ai-usage-survey/
  Einordnung: 349 Befragte, davon 87 Softwareentwickler; Gelegenheitsstichprobe und Selbsteinschätzung, also genau die Art Messung, die im ersten Experiment danebenlag.

Was es kostet

- **Aussage:** Eine wiederholte Befragung (Oktober 2024 und April 2025) zeigt eine Verschiebung vom Schreiben zum Prüfen; der Anteil derer, deren Arbeitserleben sich in mindestens einem Bereich verschlechtert hatte, stieg von 14 % auf 27 %.
  Quelle: Vella, A. & Blincoe, K. (2026): The Impact of AI Coding Assistants on Software Engineering: A Longitudinal Study. Preprint.
  Link: https://arxiv.org/abs/2605.23135
  Einordnung: Preprint, 95 wiederholt Befragte, Selbstauskunft. Präzisierung zur Folge: Die 14 % auf 27 % betreffen eine verschlechterte Developer Experience in mindestens einer Dimension, nicht eine insgesamt negative Bewertung der KI; die wahrgenommene Produktivität blieb stabil.
- **Aussage:** In einer Laborstudie waren Teilnehmende mit KI schneller (22 % weniger Zeit) und empfanden die Aufgaben als weniger belastend; trotzdem stiegen Stress und Ermüdung über die Aufgaben; der Prüfaufwand hing statistisch damit zusammen (Mediation, kein Ursachennachweis).
  Quelle: Fan, G.-R., Liu, D.-D., Pan, L.-H. & Zhang, R. (2026): When Help Hurts: Verification Load and Fatigue with AI Coding Assistants. CHI 2026.
  Link: https://doi.org/10.1145/3772318.3791176
  Einordnung: 60 Personen, drei kurze Python-Aufgaben im Labor; sagt nichts über ganze Arbeitstage. Der Zusammenhang über den Prüfaufwand ist eine statistische Vermittlung, kein gesicherter Mechanismus.
- **Aussage:** Wer eine unbekannte Programmbibliothek mit KI lernte, kam im anschließenden Verständnistest im Schnitt auf 50 %, ohne KI auf 67 %; wer die KI für Fragen und Erklärungen nutzte, verstand mehr.
  Quelle: Shen, J. H. & Tamkin, A. (Anthropic, 2026): How AI Impacts Skill Formation. Preprint.

Link: https://arxiv.org/abs/2601.20245 (Zusammenfassung: https://www.anthropic.com/research/AI-assistance-coding-skills)
Einordnung: Randomisiertes Experiment mit 52 überwiegend Junior-Entwicklern ohne Trio-Erfahrung; Test direkt danach, keine Aussage über langfristiges Lernen. Preprint eines KI-Herstellers.

Werkzeug-Tipp

Diese Folge hat keinen Werkzeug-Tipp. Der Vorschlag zum Ausprobieren braucht nur Stift oder Notizdatei: vor einer Aufgabe schätzen, wie lange sie mit KI dauert, danach die echte Zeit notieren (mit Prompten, Warten, Prüfen und Nachbessern). Wie man echte Zeiten aus Git und GitHub abliest, zeigt Folge 2.

Zum Weiterlesen

- Maier, S., Gunzenhäuser, M., Schweisthal, J., Schneider, M. & Feuerriegel, S. (2026): A meta-analysis of the effect of generative AI on productivity and learning in programming. Preprint, 23 Studien. Moderater positiver Effekt auf Produktivität mit großer Streuung, kein gesicherter Effekt auf Lernen. https://arxiv.org/abs/2605.04779
- Grundlage der Serie: Healthy AI First Software Development, Fassung 2 (01.10.2026), Kapitel 1 und 2.

Hinweis

Healthy AI First ist KI-recherchiert und KI-generiert, gesprochen von einer KI-Version der Stimme von Björn und von ihm geprüft.

Daniel und die Projektleiterin in dieser Folge sind fiktive Figuren zur Veranschaulichung. Das gilt in der ganzen Serie auch für Miriam, Jonas und Hanna.