

Stop Using a Sledgehammer

Matching AI autonomy to the task at hand

July 2026

Jess Wolfe · Customer Success, Swarmia



***Everyone's adopted the tools.
Almost no one has adopted a theory.***

60%

of work has AI integrated into it

80–100%

of tasks still need human oversight



Source: Anthropic Agentic Coding Trends Report

Two ways to misuse AI



Sledgehammer at a thumbtack

Firing up an autonomous agent to fix a typo.

High autonomy, low-stakes task.



Wrench on a screw

Tab-completing through a refactor an agent could have shipped overnight.

Low autonomy, high-leverage task.



JESS WOLFE

Customer Success · Swarmia

Hi, I'm Jess 🖐️

Customer Success Manager at Swarmia.

WHAT IS SWARMIA?

Engineering intelligence platform.

Engineering teams use Swarmia to connect GitHub, Jira, GitLab and AI tooling in one place — so leaders see who's using which AI tool, on which projects, with which results.

AI impact

track adoption & ROI

Investment balance

& capex / opex automation

DORA + SPACE

engineering performance

The five levels of AI autonomy



Miikka Holkeri, Swarmia — March 2026

L1		Assistive <i>GitHub Copilot · Cursor Tab · Grammarly · Gmail Smart Compose</i>	Autocomplete. You drive.
L2		Conversational <i>ChatGPT · Claude · Microsoft Copilot · Google Gemini</i>	Multi-context chat. You steer; model proposes.
L3		Task agent <i>Claude Code · ChatGPT Agent · Microsoft Copilot agents · Salesforce Agentforce · Cursor background agents</i>	Hand off a defined task. Get back a reviewable result.
L4		Autonomous teammate <i>Dependabot · Salesforce Einstein · Claude CoWork · Microsoft Copilot scheduled agents · Zapier AI triggers</i>	Work shows up without you asking.
L5		Agentic avalanche <i>Claude sub-agents · AutoGen · CrewAI · LangGraph · custom workflows</i>	Many agents in parallel.

INDUSTRY INSIGHT

<2% of mid-to-large orgs operate above L3. Most tools span L1–L3 — the level is set by how you use them.

Higher is not always better. · Clarity of done is your Definition of Ready for AI.

Picking a level is a staffing decision.

Domain expert on intern work

→ *wasted judgment.*

Middle-schooler running the company

→ *disaster.*

The right person on the right task

→ *leverage.*

AI is the same. The question isn't how autonomous can this be — it's how autonomous should it be?

EXAMPLE A

The vague spec

THE TASK

Improve the signup flow conversion rate.

L1



Assistive

L2



Conversational

L3



Task agent

L4



Autonomous teammate

L5



Agentic avalanche



What level is this? Drop your guess in chat — L1 / L2 / L3 / L4 / L5

EXAMPLE A

The vague spec

L2

CONVERSATIONAL

THE TASK

Improve the signup flow conversion rate.

THE RULE

If the spec fits on a sticky note and could mean three different things, you are at Level 2.



INDUSTRY INSIGHT

~70% of these are still being worked at L1 — engineers tab-completing through ambiguous specs.

EXAMPLE B

Release notes

THE TASK

You need customer-facing release notes for the latest release. Group by feature area, classify each as new / improved / fixed, skip anything tagged internal-only, use our standard format. Output: ready for editorial pass.

L1  Assistive

L2  Conversational

L3  Task agent

L4  Autonomous teammate

L5  Agentic avalanche



What level is this? Drop your guess in chat — L1 / L2 / L3 / L4 / L5

EXAMPLE B

Release notes

L3

TASK AGENT

THE TASK

You need customer-facing release notes for the latest release. Group by feature area, classify each as new / improved / fixed, skip anything tagged internal-only, use our standard format. Output: ready for editorial pass.

THE RULE

If you can describe done in one sentence and a reviewer can verify it in a single pass - promote it.

**INDUSTRY INSIGHT**

This is the level most teams under-use. The L3 adoption gap - work that should be L3 but is not - is where most AI leverage hides.

EXAMPLE C

Support ticket triage

THE TASK

A customer submits a support ticket. The agent reads the ticket, checks account tier and recent activity, classifies urgency, drafts a first response from the knowledge base, and escalates anything high-risk or low-confidence. No one manually starts the workflow.

L1



Assistive

L2



Conversational

L3



Task agent

L4



Autonomous teammate

L5



Agentic avalanche



What level is this? Drop your guess in chat — L1 / L2 / L3 / L4 / L5

EXAMPLE C

Support ticket triage

L4

AUTONOMOUS

THE TASK

A customer submits a support ticket. The agent reads the ticket, checks account tier and recent activity, classifies urgency, drafts a first response from the knowledge base, and escalates anything high-risk or low-confidence. No one manually starts the workflow.

THE RULE

If the same trigger keeps happening and the acceptable response is stable, make the agent the initiator with clear escalation rules for exceptions.

**INDUSTRY INSIGHT**

Less than 2% of mid-to-large orgs operate here today. Where the leverage gap lives.

EXAMPLE D

The PM with a hunch

THE TASK

What if the dashboard opened with a plain-English weekly summary: what changed, what looks promising, and what question a manager should ask next?

L1



Assistive

L2



Conversational

L3



Task agent

L4



Autonomous teammate

L5



Agentic avalanche



What level is this? Drop your guess in chat — L1 / L2 / L3 / L4 / L5

EXAMPLE D

The PM with a hunch

L2

CONVERSATIONAL

THE TASK

What if the dashboard opened with a plain-English weekly summary: what changed, what looks promising, and what question a manager should ask next?

THE RULE

If you are still discovering what to build, stay at L2. The output is learning — often used for a sharper Level 3 task later.



INDUSTRY INSIGHT

PM-driven prototyping is the fastest-growing AI use case — and one of the most under-supported by leadership.

EXAMPLE E

Software capitalization at scale

THE TASK

4,000 pull requests and 6,000 Jira issues last quarter. Finance needs all work classified as capitalizable or operating expense. Today: a month of report-building, meetings, defensive estimates, audit pushback.

L1

 Assistive

L2

 Conversational

L3

 Task agent

L4

 Autonomous teammate

L5

 Agentic avalanche



What level is this? Drop your guess in chat — L1 / L2 / L3 / L4 / L5

EXAMPLE E

Software capitalization at scale

L5

AVALANCHE

THE TASK

4,000 pull requests and 6,000 Jira issues last quarter. Finance needs all work classified as capitalizable or operating expense. Today: a month of report-building, meetings, defensive estimates, audit pushback.

THE RULE

L5 is not about courage — it is about control. If you cannot define the policy, evidence, confidence threshold, exception path, and audit trail, you are not ready for L5.

**INDUSTRY INSIGHT**

Most companies still do this in spreadsheets and meetings. The opportunity is defensible classification, cleaner evidence, and fewer audit surprises.

Before tech, I worked in a cancer diagnostics lab.

Diagnostics exist because you can't treat what you can't see.



DIAGNOSE — Find the driver, not the symptom.

Molecular tests identify which gene is broken — for example the BCR/ABL gene fusion in chronic myeloid leukemia.

TARGET — Match the treatment to the driver.

Gleevec is the proven targeted treatment for t(9;22)(q34;q11) found in CML and some other cancers

MISS — Treat blind, and best case it fails.

Without a target for treatment, treatment itself can cause therapy-related MDS. A cure without the right target will make the system sicker

Organizations break the same way.

Organizations are living systems

When a process breaks, the symptom shows up downstream of the break.

- 1 A process breaks
- 2 The pain appears downstream
- 3 The loudest voice gets treated
- 4 A fix gets bolted on — at the wrong point
- 5 ***The fixes become part of what's broken.***

What each level removes

Every step up the ladder removes one human function.

L5

Removes per-item review — *you stop checking each result*

L4

Removes initiation — *you stop deciding when it works*

L3

Removes step-by-step steering — *you stop watching it work*

Every human function you remove must be replaced by something your organization can see.

How much AI autonomy is safe?

Your visibility into your own system of work sets the ceiling on how much autonomy you can safely deploy.

STAGE	WHAT YOU CAN SEE	CEILING
Foundational <i>"flying blind"</i>	Anecdotes. No baseline, no investment data.	L2
Crawl <i>manual visibility</i>	Tagged work. Your first real numbers on what AI is changing and how it affects your system of delivery.	L3
Walk <i>instrumented</i>	Standardized metrics on a cadence. Working agreements enforced where work happens.	L4
Run <i>self-regulating</i>	Real-time guardrails. Review by exception.	L5

Breaking the ceiling

Deploy above your ceiling and you don't get a crash. You get these.

FOUNDATIONAL + L3

**AI usage up,
throughput flat.**

*Usage is not value — tool impact
correlates with mature practices,
not adoption.*

CRAWL + L4

**The bottleneck moves
downstream.**

*Your most senior people drown in
review of plausible-looking work.*

WALK + L5

**Output faster than the
system can absorb it.**

*That's not throughput — it's a
new bottleneck with better
marketing.*

<2%

of mid-to-large organizations operate above Level 3

Not a tooling gap. A maturity gap.

The tools have been ready for a while. Visibility is the part that's rare — and visibility is buildable.

What to measure to raise your ceiling

Each failure has a metric that prevents it. Measure these, and the ceiling rises.

SO L3 PAYS OFF

Measure impact, not usage.

Cycle time and review time, before vs. after — not logins or seats.

Swarmia shows AI adoption and AI impact side by side, so you see whether the tool moved delivery or just got installed.

SO L4 IS SAFE

Watch the downstream load.

Review time, batch size, and work in progress — the early-warning signals that AI volume is drowning your reviewers.

Swarmia surfaces review and flow metrics per team.

SO L5 HOLDS

Can the system absorb it?

Measure flow efficiency at the highest work level — and define the judge and escalation path before you scale, not after.

Swarmia shows flow efficiency and investment balance.

Capture a baseline BEFORE you roll out L4/L5 AI solutions

The AI Gate & Ceiling Cheatsheet

GATE 1 — THE TASK

Can I describe what done looks like?

No → clarity problem, not an AI problem.

GATE 2 — THE SYSTEM

Can we see the system it lives in?

No → your ceiling is L2–L3.

THEN PICK THE LEVEL

Done in one sentence **L3**

Same trigger recurring **L4**

High volume + automated judge **L5**

None of these — and that's fine **L2**

THE AUTONOMY CEILING

Foundational

flying blind — anecdotes, no baseline

L2

Crawl

manual visibility — tagged work, impact being measured

L3

Walk

instrumented — standardized metrics on a cadence

L4

Run

self-regulating — review by exception

L5

Ceiling = the highest level you can safely deploy.

The bottleneck has moved... or has it?

Quality of intent

has almost always been your bottleneck

Context feeding clarity of done is your Definition of Ready for AI.

AI is an amplifier. Adopt without clarity and you scale your problems, not your output.

**Stop swinging a sledgehammer.
Match the tool to the task.**

Invest in the quality of your intent.

Know your ceiling — then raise it.

Two ways to go deeper

See your team's AI adoption — and your stage.

- AI tool adoption
- Engineering effectiveness
- Investment & ROI
- Team health & flow



Book a 15-minute call.

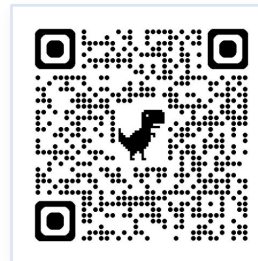
We'll show you your own data. No commitment.

THE BOOK

**Build: Elements of an Effective
Software Organization**

**3rd edition — free pre-order,
ships August 2026**

*Foreword by Nathen Harvey,
DORA Lead, Google Cloud*



Pre-order — it's free.

swarmia.com/pre-order-build

Thank you

Jess Wolfe

jess@swarmia.com

LinkedIn:

<https://www.linkedin.com/in/thejessicawolfe/>



The framework:

swarmia.com/blog/five-levels-ai-agent-autonomy

