

# My agent workbench

Making things with AI.  
Keeping enough evidence to trust them.

**John Lee**

AI & Ales · Jamestown Frontier Society  
October 15, 2026 · Billsburg Brewery



# The tools around my agent

## **Codex**

A working partner in my projects, files and terminal

## **Skills + Markdown**

Reusable instructions and decisions that survive a chat

## **Git + worktrees**

Separate experiments, inspect changes, recover a state

## **Pixi + DataLad**

Recover environments and connect data to the work

[My skill workshop](#)

[Worktree tools](#)

[Execution records: duct](#)

# Keyhatch

## Selective credential access for agents

Approve selected credentials once.  
Allow repeated retrieval within a limited window.

Native macOS app + local command-line client

Expiry stops future retrieval. It cannot recall a disclosed secret.

[Project + source](#)

[Design and limits](#)

Experimental build

# The Software Engineer Reborn

I used an interactive preparation space to work through the ideas.



One causal chain holds most of the material together.

The strongest version is not "software engineers will survive." It is an account of what becomes scarce when implementation

## Explore the preparation



# The Software Engineer Reborn

A vision of human-AI collaboration.



/ 25

## Open the presentation

# What did we just make?

## WHAT DID WE JUST MAKE?

Managing research objects with AI

John Lee

New York AI Toastmasters

← → navigate · F fullscreen · N notes

**A research object includes the result and what it takes to inspect or rerun it.**

STAMPED asks about the package: its history, its parts, its environment, and how someone else can use it.

[Research objects with AI: the talk](#)

[STAMPED principles](#)

# Orinoco Lite

## Connected research information

People, projects and outputs share explicit relationships.

An editor proposes changes.  
GitHub review preserves the history.  
The records feed a website and graph.



[Project + source](#)

[Explore the graph](#)

[Poster](#)

# STAMPED in Practice

## Skills that improve through scientific work

## Study record on Zenodo

## Next: run, assess, improve, repeat

## Compare four assistance conditions. Test quality, effort and consistency.

STAMPED - US-RSE 2026

STAMPED in Practice

Checking AI-assisted analysis and its research record



John A. Lee  
Austin MacDonald  
Cody C. Baker  
Yaroslav O. Halchenko  
Center for Open Neuroscience  
Psychological and Brain Sciences  
Dartmouth College



Poster  
+  
Source

An AI assistant can write analysis code. Can another person check its result, rerun the work, and reuse what it produced? STAMPED gives us seven practical questions to ask about that research.

We compare the same tasks with and without five reusable skills.  
A skill is a short set of instructions and checks an assistant can follow. We check scientific results and the records needed to reproduce them.

1 - WHAT WOULD MAKE THIS WORK REUSABLE?  
Seven properties, with evidence for each

Three paired tasks show useful evidence and limits. A recorded command alone does not show that an analysis is correct.

| Property       | Practical question                  | Original abstract: tool plan            | Three pairs: evidence and gaps                                                                  |
|----------------|-------------------------------------|-----------------------------------------|-------------------------------------------------------------------------------------------------|
| Self-contained | Is everything needed identified?    | DataLad + git-annex                     | Partial: both arms<br>Pinned inputs and locked software; offline setup untested.                |
| Trackable      | Can we follow how it changed?       | Git + DataLad run + conda + NDM + PROCV | Partial: both arms<br>Git and raw audits retained; native DataLad rerun incomplete.             |
| Actionable     | Can someone run or use it?          | Tested BIOS Apps + Pto tasks            | Demonstrated: both arms<br>All six science commands replayed; full reports untested.            |
| Modular        | Can a part be reused on its own?    | DataLad composition                     | Demonstrated: both arms<br>Three supplied datasets; workers separated code and outputs.         |
| Portable       | Can it run in another environment?  | Apptainer                               | Partial: both arms<br>New software environments; one Mac; other systems untested.               |
| Ephemeral      | Can a fresh workspace reproduce it? | Apptainer + BABS + Slurm                | Partial: both arms<br>Outputs rebuilt in new checkpoints; shared caches retained.               |
| Distributable  | Can others obtain the right parts?  | Persistent git-annex remotes            | Partial: both arms<br>Input format known; new code licensing and external retrieval unverified. |

Tools preserve the original proposal; many were not used. Judgments cover these three pairs. The coordinator supplied datasets and ran independent replay; later log packaging is curatorial. Containers and clusters are untested.



Study + evidence

2 - A SMALL, INCREASINGLY DEMANDING CASE STUDY  
How do response times change across runs?

We use public mixed-gametes task records from OpenNeuro [doi:000005](#): 16 participants and three runs each. We summarize recorded decision times; we do not enter any they changed.

|                 | 01<br>One participant<br>Compare the median response time in the first and third runs. | 02<br>The group<br>Estimate the average within-person change and its uncertainty. | 03<br>Sensitivity<br>Vary the summary and response-time rule; add the middle run. |
|-----------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| One participant | 278 s                                                                                  | 358 s                                                                             | 413 s                                                                             |
| Group           | 387 s                                                                                  | 474 s                                                                             | 567 s                                                                             |
| Sensitivity     | 383 s                                                                                  | 474 s                                                                             | 567 s                                                                             |

Each stage uses two fresh assistant contexts. Progress depends on a correct result and an independent successful replay of the preceding stage.

Checked scientific results  
One participant: median response time 1272 ± 1261 ms; change -11 ms. Both arms agree.  
16 participants: average within-person change +31.3 ms; 95% bootstrap interval [70.3, 132.7] ms. Both arms agree.  
Change is third minus first run; the interval includes zero. Missing responses limit interpretation. This is not a causal effect.  
Sensitivity: median/mean summaries, with or without a 0.2-0.3 s response-time band, gave +11.2 to +58.7 ms; all four 95% intervals include zero.

Scope change: public behavioral data replace the proposed ABCD neuroimaging case. The ABCD result has not been reproduced.

3 - WHAT DO THE SKILLS ADD?  
A matched comparison

Both arms receive the same data, task, software, STAMPED references, and shared account controls. Only the treatment arm is instructed to read the five new skills.

| Baseline                 | With skills               |
|--------------------------|---------------------------|
| Common task and controls | Common task + five skills |

Each attempt: 8 minutes - git & extra - ultra reasoning - fresh tool workspace.

Reusable research guidance  
The skills turn the seven properties into steps: define the research object, preserve evidence, make work runnable, check it, and prepare it for reuse.



Read the skills

REPORT BENEFITS AND COSTS TOGETHER  
What do we check?  
• Data checks: separate numerical and row-level checks.  
• Replay: remove outputs, then rebuild in a fresh checkout.  
• Costs and gaps: time, tokens, setup, interventions, and failures.

ONE ATTEMPT PER ARM AND TASK  
Checked outcomes

| Task        | Baseline                                                            | With skills                                                         |
|-------------|---------------------------------------------------------------------|---------------------------------------------------------------------|
| One person  | 278 s<br>Data checks: Passed<br>Replay: Passed                      | 358 s<br>Data checks: Passed<br>Replay: Passed                      |
| Cohort      | 387 s<br>Data checks: Passed<br>Replay: Passed                      | 413 s<br>Data checks: Passed<br>Replay: Passed                      |
| Sensitivity | 383 s<br>Data checks: Passed<br>Replay: Passed<br>Screening: Passed | 474 s<br>Data checks: Passed<br>Replay: Passed<br>Screening: Failed |

Worker time only; setup and evaluator replay are separate. Replay used a fresh checkout on the same host.

Screening and integration failures  
The skills arm in task 03 failed the required case-insensitive "NA" check. Those failures were absent from the real data. Missing validation software and native DataLad rerun also caused failures.

Limits of this study  
All trials share an account, host, and software caches. Fresh contexts do not isolate account settings.  
An unblinded study-design agent checked the outputs; the authors develop STAMPED and these skills. This small case series cannot establish general effectiveness.

Three local pairs, October 2026. Source: Tom et al. (2007), mixed-gametes task, OpenNeuro doi:000005 (PDDL). Code, input-versions, protocol, and review notes are linked from the study.  
STAMPED mirrors Self-contained, Trackable, Actionable, Modular, Portable, Ephemeral, Distributable. Poster layout adapted from CON's US-RSE 2026 OERINOCO Live poster.  
OpenAI Codes assisted the analyses and poster draft.



Explore CON



Dartmouth College  
Psychological and  
Brain Sciences



Supported by  
Repligen

Research workshop in development. Transfer remains to be tested.

[Poster PR #35](#)

[Study on Zenodo](#)

[Next repository](#)

# What would convince you?

What evidence do you need before you trust something an agent made?

I'm exploring how to make that evidence easier to keep and inspect.

[leej3.github.io/ai-and-ales-2026](https://leej3.github.io/ai-and-ales-2026)

[All links, downloads and editable source](#)