

— TWODOT ENGINEERING · HACK DAY ATTEMPT

BAS Agent

What two people and a swarm of agents tried, what the experiment taught us, and where BAS automation ends—and trustworthy assistance begins.

HACK DAY RETROSPECTIVE · 2026-07-17 · WORK IN PROGRESS

— THE BRIEF

A recurring compliance workflow, drawn as **five stages**

27

requirements across the end-to-end flow

5

sequential stages, from calendar to completion

3

integration surfaces: Xero, Google Chat, scheduler

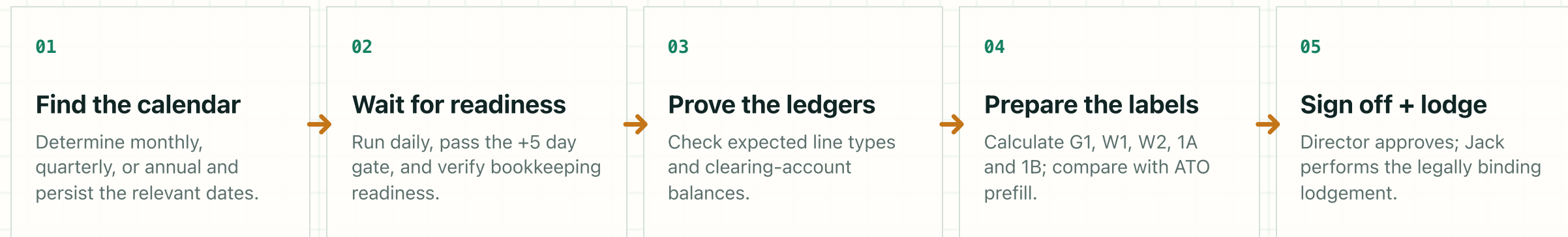
2

humans remain accountable: Jack and the director

Grounded in PROJECT-052 and the Stage 1–5 process diagrams.

— TARGET WORKFLOW

A funnel of increasingly expensive checks



Run the same goal through multiple agent setups

Nx

Rather than betting on one decomposition, we deliberately sent multiple agents at the whole problem and compared what stuck.

Redundancy was the experiment: keep the goal fixed, vary the working environment, and preserve multiple independent paths to a result.

— TWO PEOPLE, TWO TRACKS

Same outcome; different agent strategies

Callum → Hermes Agent

Callum used Hermes with Codex SOL to build the end-to-end solution that now runs locally.

Craig → multiple clankers

Craig built github.com/twodotai/exp-bas as a shared context mono repo, then launched multiple attempts. The result running in dev came from Claude Fable and direct Codex SOL.

The comparison was part delivery strategy, part learning experiment.

HACK-DAY OUTCOME

We both shipped

1

Callum: local and working

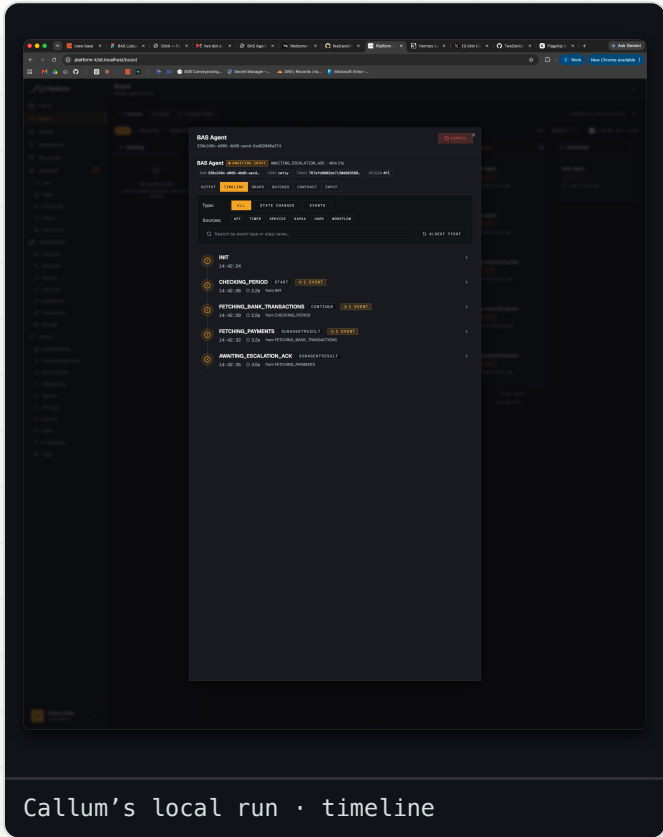
Hermes did the build, with Codex SOL. The resulting BAS Agent is running locally and working.

2

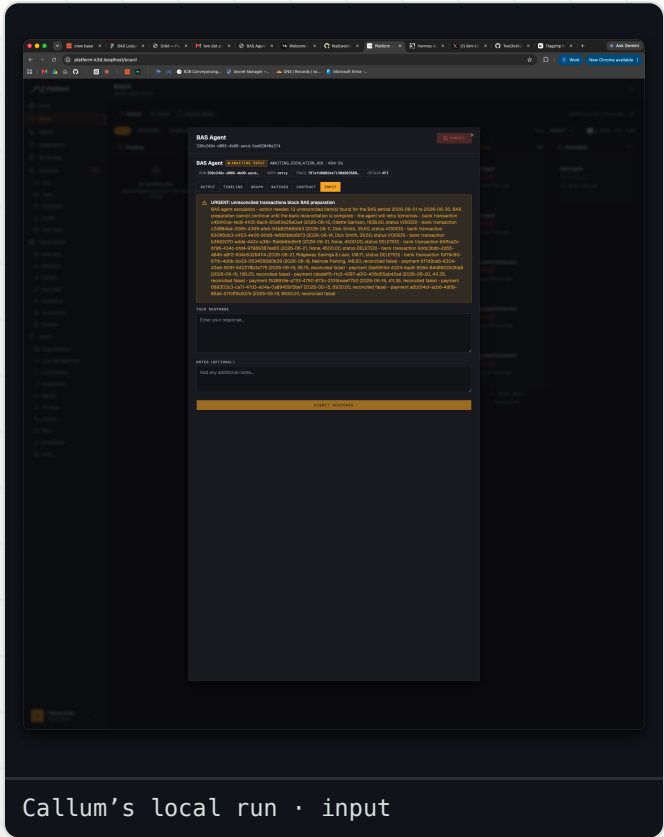
Craig: running in dev

Claude Fable and direct Codex SOL produced the Craig-led implementation now running in dev.

* The dev build may or may not be slop. It is, however, running.



Callum's local run · timeline



Callum's local run · input

Package the problem as a reusable context mono repo

01

Read Crewbase

Use the existing read script to feed Jack's written tasks into a clanker.

02

Verify the summary

Publish a clear brief at bas-agent.daybreak.exp.twodot.ai and check it with Jack.

03

Find the missing context

Jack pointed to a FigJam diagram that otherwise would have been missed.

04

Commit the context

Rewrite the FigJam boards as Markdown + Mermaid in the exp-bas mono repo.

Craig's sessions used the same verified brief, reconstructed process, API map and build plan.

Shared context mono repo: github.com/twodotai/exp-bas

API-MAP.md found the missing integration layer

Four OAuth scopes were missing

`accounting.reports.trialbalance.read`

`accounting.reports.taxreports.read`

`accounting.reports.profitandloss.read`

`payroll.payruns.read`

Four gateway reads were missing

Trial Balance, Profit & Loss, published BAS reports and pay runs needed typed, centrally authorised capabilities. Empty BAS results must preserve the human path.

Consent + infrastructure still mattered

Existing Xero integrations need re-consent. Google Chat send, reply and `call_human` already existed; its production Kafka topic was a separate infra dependency.

Journals was an expensive non-dependency

The only Advanced-tier, security-assessment path could be avoided: Trial Balance plus transaction endpoints supply the evidence. BAS write and lodgement still have no API.

— THREE CRAIG-LED EXPERIMENTS

Vary the environment, not the goal

01 • Mono-aware

Ran against mono using Will's instructions to bring up a local environment and work within the platform.

SHIPPED TO DEV

02 • Raw attempt

Knew nothing about mono and was asked to solve the problem independently, purely to expose what it would invent or miss.

LEARNING RUN

03 • Gas City

Noticed prior work already existed, then asked permission to stop, wait for the other attempt, and improve it instead of duplicating it.

SELF-COORDINATED

The Mono-aware implementation shipped to dev via Claude Fable + direct Codex SOL. The Gas City run also surfaced coordination behaviour we did not explicitly ask for.

— RATE LIMITING BECAME PRODUCT WORK

Rate limiting forced a provider switch

Claude

Rate-limited while several sessions were still active.



Codex

Craig wrote a provider-switching tool and moved the sessions across so they could keep running.

Provider portability went from “nice to have” to critical path.

— THE SWARM WAS BIGGER THAN OUR TEAM

Agent telemetry made the company experiment visible

3

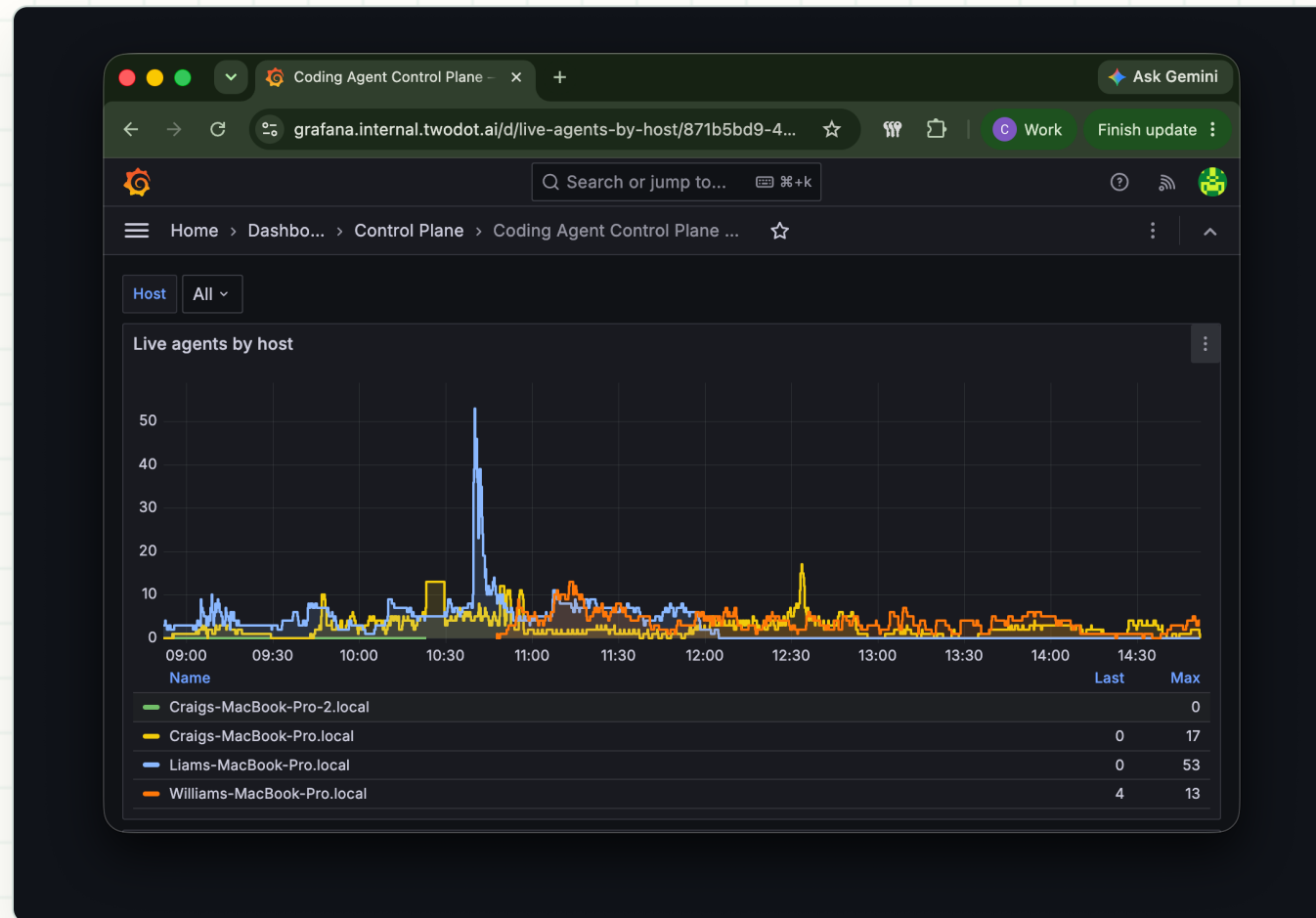
Three hackathon participants

Each was sending lifecycle hooks into the shared agent-metrics collector.

≈

Liam hit 50 concurrent agents

That put one participant halfway to Grant's goal of 100 concurrent agents.



The requirements described the **Xero UI**, not the Xero API

What the flow says

- ↳ Open Reports → General Ledger
- ↳ Read the draft Activity Statement
- ↳ Check the ATO-prefilled figures
- ↳ Enter W1 and check lodgement status



What the API exposes

- ↳ No General Ledger report endpoint
- ↳ No draft Activity Statement or ATO prefill
- ↳ Report endpoints are read-only
- ↳ No BAS lodgement endpoint

The method had to change while preserving the business outcome.

Automate the evidence; make the boundary explicit

Automate

Calendar lookup, date gates, Trial Balance checks, transaction aggregation, payroll traversal, label calculations, audit trail and escalation packets.

Human attests

Bank-reconciliation completeness, ATO-prefill comparison, PAYGI, director approval, and any accounting judgement the API cannot prove.

Human executes

Jack enters the required statement values and clicks Lodge. The public API cannot perform that irreversible action.

A BAS preparation and review agent is buildable. A BAS lodgement agent is not.

— THE HARDEST LABEL

W1 ≠ PayRun . Wages

The same gross wages can produce a different W1 because reporting is configured per pay item.

PayItems

Read which earnings are reportable as W1 and which deductions are exempt.

PayRuns

Keep posted runs whose payment date falls inside the BAS period.

Payslips

Walk each employee’s earnings and deduction lines—the expensive N+1 step.

Cross-check

Tie computed W1 and W2 back to independent Trial Balance movement or stop.

— THREE HONEST STOPS

A false green is worse than a hand-off

1

Bank reconciliation

The API can find some unreconciled transactions, but cannot prove that no unseen bank-feed lines remain. Jack must attest completeness.

2

ATO prefill

The draft prefilled Activity Statement is not available. The agent supplies its figures; a human compares and enters them.

3

Lodgement

No Xero API can lodge a BAS. The designed outcome is a verified package and a deep link for Jack—not a pretend automation.

— SAFETY ARCHITECTURE

Keep the LLM away from the money—and production

Demo Company only

Discover the demo tenant, block the production tenant and denylist its name. Ambiguity fails closed.

GET-only transport

Enforce read-only requests immediately before the network. The model never receives tenant, header, or write controls.

Deterministic calculation

Pure decimal arithmetic produces the labels and provenance. An LLM may narrate an escalation; it never touches a number.

Every unknown classification, failed cross-check, or safety mismatch becomes a loud stop with evidence.

— WHAT THE ATTEMPT PRODUCED

An implementation-grade map—and two working builds

Process design

Five diagrams reconciled into one narrative with branches, loops and escalation semantics.

API-MAP .md

Mapped requirements to APIs and scopes, then produced the platform integration backlog implemented through mono PR #3762.

Corrections

Three plausible W1 designs were rejected as the team learned what the source data actually means.

Build plan

Pure calculation first, then authentication and safety, sources and cross-checks, state, chat, and a TypeScript parity port.

Context mono repo: github.com/twodotai/exp-bas · [requirements.txt](#) · [BAS-AGENT.md](#) · [API-MAP.md](#) · [PLAN.md](#) · [Stage1-5.png](#). Outcome: Callum's build works locally; Craig's build is running in dev.

— WHAT STILL NEEDS A RULING

The questions that can change a lodged number

How are BAS payments booked?

The ATO-line classifier depends on whether settlements carry a reliable contact and source pattern.

How does salary sacrifice affect W1?

Xero's aggregates do not document enough; this needs empirical reconciliation with a representative organisation.

Which basis and registrations apply?

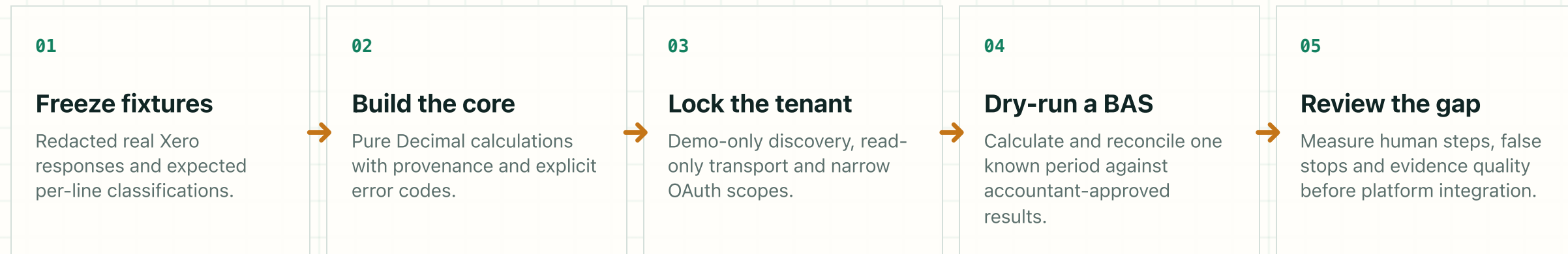
Cash versus accrual changes GST calculations, while frequency alone cannot safely determine the labels.

What happens to annual clients?

Annual appears in the first-stage requirement, but the downstream flow only defines monthly and quarterly branches.

— A CREDIBLE NEXT SLICE

Prove one period, end to end, without lodging



— DELIVERY SUMMARY · CAPTURED 2026-07-17

A working slice shipped; the production edges stayed explicit

Shipped and running

Callum's local build, Craig's dev build, the 27-task BAS workflow, Kylie exposure, the presentation publisher, and the context repo with its implementation-grade API map.

Built and in review

Xero report/pay-run scopes and reads, seven workflow-hardening PRs, the production Chat topic, and the per-host agent telemetry dashboard.

Did not ship

Direct ATO-prefill access, BAS write/lodgement, a production journal robot, or proof that the full workflow handles accountant-approved real responses end to end.

EXPERIMENT READY: LedgerPilot can now test review-first Xero journal UI automation in a demo organisation.

A browser robot is a probe—not a shortcut around the API

UI

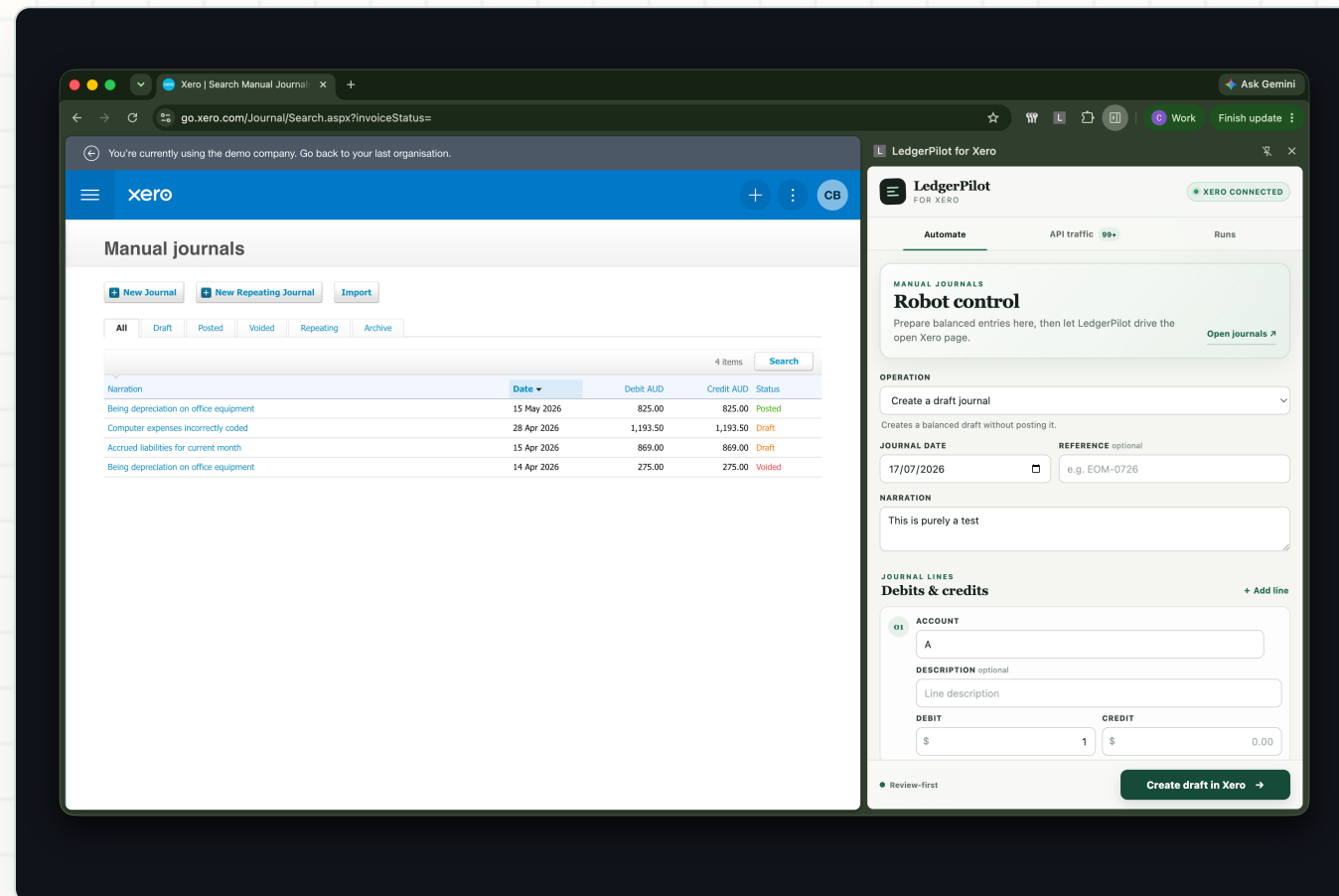
Review-first Manual Journals automation

The Chrome extension can drive create, update, copy, post, reverse, delete and void through Xero's visible UI. Risky final actions require confirmation.



Useful elsewhere; not required for BAS

BAS evidence can use Trial Balance plus transactions without Journals. Keep this experiment for Xero work where the usable API surface is missing or gated.



Xero documents a Manual Journals API; this experiment tests the signed-in UI without an OAuth client and remains subject to Xero permissions, validation and UI changes.

— LESSONS LEARNED

What we would carry into the next build

Redundancy can beat decomposition

Parallel full-problem attempts created multiple independent paths to a usable result.

Share verified context

A verified brief and machine-readable diagrams gave independent agents the same domain model and API constraints.

Ask tools to explain themselves

Hermes could guide its own setup, while the configuration UX still needs testing with actual domain experts.

Vary the harness

Mono-aware, repo-naïve, and Gas City runs exposed different assumptions and coordination behaviours.

Design for provider failure

Rate limits can interrupt active runs. A provider switch should be routine, not emergency tooling.

Instrument the swarm

Shared hooks turned invisible agent activity into a company-level signal we could inspect and discuss.

— HACK-DAY TAKEAWAY

Don't automate the click. Automate the confidence.

The opportunity is a tireless analyst that proves what it can, shows its working, and arrives at the human boundary with a recommendation—not a bot that pretends the boundary is not there.

NEXT: replace time-stamped statuses with final results · prove one BAS period · decide whether the journal robot merits hardening

Craig's BAS Agent pull requests

PR	STATUS	WHAT IT IS
mono #3749	OPEN	Per-host agent telemetry used in this deck, plus a bounded CACP row-window OOM fix.
mono #3750	MERGED	Initial 27-task, five-stage BAS workflow with unit and fake-integration E2E coverage.
mono #3760	MERGED	Initial BAS hack-day deck and the engineering-presentation CDN publisher.
mono #3762	OPEN	Xero report/pay-run scopes and typed gateway reads discovered through API-MAP.md.
mono #3763	MERGED	Exposes the BAS workflow to Kylie-enabled organisations and the assistant catalogue.
mono #3764	MERGED	Expands the deck with the team's agent strategy, telemetry and lessons learned.
mono #3767	MERGED	Embeds the Grafana agent-concurrency capture in the presentation.
mono #3769	MERGED	Adds the context-repo and API-MAP.md findings plus this linked PR appendix.
mono #3772	MERGED	Marks the mono-aware BAS implementation as shipped to dev.
mono #3773	OPEN	Adds the PDF export and the higher-resolution Callum screenshots.
infra #1653	OPEN	Provisions the production Kafka topic required by the Google Chat subagent.

Callum's BAS Agent pull requests

PR	STATUS	WHAT IT IS
mono #3753	OPEN	Propagates Xero reconciliation evidence and escalates false or missing status.
mono #3755	OPEN	Audits operator-approved prefill adjustments with responder and timestamp provenance.
mono #3756	OPEN	Refreshes BAS periods and handles extensions, weekends and configured public holidays.
mono #3757	OPEN	Persists reconciliation GLs and calculates period-scoped labels with ATO exclusion.
mono #3761	OPEN	Routes director sign-off through a dedicated Google Chat space.
mono #3765	OPEN	Gates preparation by inferred lodgement frequency and preserves Xero line evidence.
mono #3766	OPEN	Covers the daily scheduler E2E and repairs canonical browser-session test setup.
mono #3770	OPEN	Persists evidence and reasons for prior-period reconciliation decisions.
mono #3771	OPEN	Proves matching and corrected planning paths both converge at Fill in W1.

State captured from GitHub on 2026-07-17.