CDAO OPENING KEYNOTE

AI Agents Unleashed

Why Most Fail and How Some Deliver Real Impact

- Moving beyond pilots to real-world deployment
- Augmenting human judgment, not replacing it
- Trust — not technology — as the real barrier
- Turning agents into measurable business value

A MOMENT THAT CHANGED HOW WE THINK ABOUT AI

A physical therapist. 7 PM. Still typing.

“

One of our physical therapists finishes seeing patients at 4 PM. But the day doesn't end there. Documentation — visit notes, plans of care, progress summaries — takes another two to three hours. Three hours that can't be spent with patients. Three hours of burnout, one shift at a time.

When we started looking at AI agents, this is the problem we kept coming back to. Not because it was the easiest to solve - it wasn't. But because the stakes were real: fewer patients seen, worse outcomes, clinicians leaving the profession.

The question we asked ourselves:

"Can an AI give clinicians their time back — and actually improve what happens for patients?"

That question led us to three agents.

WHERE WE'RE HEADED

Today's roadmap

01 

The State of Agents

Why most programs stall before they scale

02 

The Design Principle

Augment judgment, don't replace it

03 

Three Agents, Real Results

What we built and what we measured

04 

Trust as the Barrier

Data quality, hallucination, compliance

05 

Turning Agents into Value

The framework behind measurable ROI

Everyone has a pilot. Almost no one has scale.

42%

of companies abandoned most AI initiatives before reaching scale

~5%

of enterprise GenAI pilots reach production with measurable ROI

~2 in 3

of organizations cite security and risk concerns — not capability — as the top barrier to scaling agentic AI

Sources: S&P Global Market Intelligence, 2025 Enterprise AI Survey; MIT NANDA, State of AI in Business 2025; McKinsey, 2026 AI Trust Maturity Survey.

Mentimeter

Poll #1

Where are your AI Agents today?

- Still exploring / no agents in production
- One or two pilots running
- Live in production, limited scope
- Scaled across multiple functions

menti.com

8854 8567

Waiting for participants

THE PATTERN BEHIND THE FAILURE RATE

Why most agent pilots never scale

You've probably seen a version of this list before — here's what we actually did about each one.

Built for the demo, not the workflow

Tuned to impress in a boardroom, not to handle messy edge cases in production.

No owner when the pilot ends

Runs in an innovation lab with no handoff to the team that owns the process.

Data was never production-ready

Stale, fragmented, or contradictory knowledge — the agent inherits the mess.

Success was never defined

'It seems to work' replaces a hard metric. Nothing to prove ROI against.

THE DESIGN PRINCIPLE THAT SEPARATES WINNERS

The best agents augment judgment. They don't replace it.

Replace-and-hope

- Removes the human from the loop entirely
- Optimizes for fewer headcount, fast
- Fails silently on edge cases
- Trust collapses the first time it's wrong

Augment-and-escalate

- Resolves the routine, escalates the complex
- Gives people time back for judgment calls
- Knows the boundary of its own confidence
- Trust compounds — adoption grows over time

FROM OUR OWN DEPLOYMENT

Three agents. Three real outcomes.

Not a roadmap. Production deployments with measured results — including one that stretches the definition of "agentic," which we'll address head-on.

ServiceNow Resolver

88% of tickets resolved or AI-assisted

Knowledge Assistant

40% reduction in information-seeking tickets

Ambient AI Solution

More patients per week, better outcomes

CASE STUDY 01 — SERVICENOW RESOLVER

88% of L1 tickets resolved or AI-assisted

Phase 1
Monitoring Only

Agent responds. Humans review every response before action. Accuracy logged, no auto-resolution.

Phase 2
Tune & Validate

Edge cases identified and addressed. Confidence thresholds set. Escalation paths defined.

Phase 3
Auto-Resolve + Monitor

Agent resolves autonomously within confidence bounds. Ongoing monitoring continues.

49%

Fully
auto-resolved

39%

AI-assisted
human finishes

12%

Needs full
support team

CASE STUDY 02 — AGENTIC KNOWLEDGE ASSISTANT

40% reduction in information-seeking tickets

How it works

- Indexes policies, wikis, SOPs, and prior resolved tickets
- Retrieves the most relevant source before answering — grounded, not generative
- Cites the source so front desk can verify the answer themselves
- Flags low-confidence answers rather than guessing
- Escalates to a human when confidence falls below threshold

40%

fewer tickets submitted because employees got answers instantly

- Cites sources, not just answers
- Says 'I don't know' when it doesn't
- Routes to a human when uncertain

CASE STUDY 03 — AMBIENT AI SOLUTION

More patients seen. Better outcomes. Less clinician burnout.

What the agent does

- Listens to patient-clinician conversations in real time
- Auto-generates structured visit notes and care summaries
- Surfaces relevant clinical context during the encounter
- Clinician reviews and approves — always in the loop

What we measured

↑ Patients per week

Clinicians can see more patients because documentation no longer eats their schedule.

↑ Patient outcomes

More time with patients, more accurate notes = better continuity of care.

↓ Clinician burnout

End-of-day documentation burden significantly reduced.

As promised: is this "truly" agentic? It perceives, reasons, and acts within a workflow — we call that agentic in the sense that matters, autonomous value with a human still in the loop.

The one we paused before it shipped

01
What We Built

An agent to answer payor billing questions instantly — referral rules, filing limits, visit caps — replacing manual search through thousands of hand-maintained payor files.

02
Why We Stopped

No shared schema across payor files, decision-critical policies locked inside scanned images, and the same rule reading differently contract to contract — accuracy risk we couldn't accept on denial-driving decisions.

03
What's Next

Moving the agent onto Claude Sonnet — strong enough to read the scanned pages and inconsistent layouts directly — and re-piloting with mandatory source citation on every answer.

Paused

before
go-live

3 root

causes
identified

Re-pilot

in
progress

We couldn't load the Menti slide

Exit slideshow mode, then select the Menti slide you want to add. After selecting your slide, start the slideshow again.

If the issue persists, please reach out to us at hello@mentimeter.com

Copy diagnostics

THE REAL OBSTACLE ISN'T CAPABILITY

Trust is the barrier — not technology

Data quality

Agents are only as good as what they retrieve. Fragmented, stale, or contradictory sources produce confident, wrong answers. We fixed it with a data-ownership model — every source has a named owner accountable for its accuracy.

Hallucination

Without grounding and calibrated confidence, agents fabricate plausible-sounding answers — the fastest way to lose a user forever.

Compliance

Every autonomous action needs an audit trail, access boundaries, and a clear escalation path for regulated decisions.

WHAT WE DID ABOUT IT

Earning trust, deliberately

Start in monitoring-only mode	Don't let the agent act until you've validated accuracy on real production data.
Ground every answer in a source	Both agents retrieve before they respond, and cite what they used.
Calibrate confidence, not just accuracy	Low-confidence cases escalate to a human rather than guess.
Log every autonomous action	Reasoning and resolution steps recorded for audit and continuous improvement.
Keep a human checkpoint on risk	Anything touching access, identity, or patient data requires a review path.

We couldn't load the Menti slide

Exit slideshow mode, then select the Menti slide you want to add. After selecting your slide, start the slideshow again.

If the issue persists, please reach out to us at hello@mentimeter.com

Copy diagnostics

FROM IMPRESSIVE TO VALUABLE

Turning agents into measurable business value

1. **Anchor to a KPI that already exists**

   Tie the agent to a metric leadership already tracks — ticket volume, AHT, CSAT, patients per week. Not a new AI-specific vanity metric.

2. **Baseline before you deploy**

   Measure the current state honestly. You can't claim impact without a 'before.'

3. **Monitor the agent, not just the outcome**

   Track resolution rate, escalation rate, and confidence — not just usage counts.

4. **Report in business language**

   '40% fewer information tickets' lands with leadership. '94% accuracy score' doesn't.

HOW WE RESOURCED THIS

No new budget. No new headcount.

Started with unused licenses

We were already paying for AI tooling nobody was using — we redirected it instead of asking for new budget.

Built by our own developers

No outside vendor build. Our existing engineers stood up the first agents alongside their regular work.

Quick wins earned the room

Small, visible results — not a roadmap — were what got leadership's attention and unlocked real support.

Now scaling through champions

Champions build inside clear guardrails and governance. Ownership shifts from IT to the business areas each agent serves — accountable for outcomes, not just uptime.

Buy first. Build only when it's worth owning.

Buy
If it exists, license it

A strong off-the-shelf solution beats a custom build. Ambient AI — bought, not built.

Build — low-code
If not, build fast

Low-code/no-code tools business teams can own. ServiceNow Resolver — Copilot Studio, Power Automate.

Build — custom
Only when it's earned

When requirements demand real control. Payor Agent — built in Claude Code, running on Sonnet inside Azure AI Foundry, with AI Search as the vector store.

Buy

when a good
option exists

Low-code

decentralized —
champions build

Custom

centralized —
one dev team

WHERE TO START MONDAY MORNING

Your first 90 days with agents

Days 1–30

Pick one high-volume, well-documented pain point. Baseline the current state. Agree on the KPI before you build.

Days 31–60

Deploy in monitoring-only mode. Review every output with a human. Tune on real production data, not a test set.

Days 61–90

Enable autonomous action in the highest-confidence categories. Monitor. Report the business KPI. Then expand.

The agents that scale aren't the ones with the best demo — they're the ones earning trust one decision at a time.

Thank you

Let's talk about turning your agents into measurable impact.

Questions • Discussion • Connect

– Start with monitoring — earn the right to auto-resolve

– Trust is built by saying 'I don't know' as much as by getting it right

– The metric that matters is business impact, not model accuracy