Danny Wheller, VP of Business & Strategy at Hebbia, on vertical vs horizontal enterprise AI

TEAM

Jan-Erik Asplund
Co-Founder
jan@sacra.com

Published on **Jun 30th, 2025**

Danny Wheller, VP of Business & Strategy at Hebbia, on vertical vs horizontal enterprise AI

By **Marcelo Ballvé**

Background

After we published our report on Hebbia last fall, we got the opportunity to speak with Danny Wheller, VP of Business & Strategy at Hebbia. Hebbia is a top enterprise AI player that last raised a $130M Series B led by Andreessen Horowitz at a $700M valuation.

Key points from our conversation via Sacra AI:

- **Versus horizontal enterprise AI like Glean—pricing at $20/mo per seat, serving mid-market & enterprise and going after wall-to-wall adoption in the org—Hebbia positions itself as a deeper vertical solution for finance & legal in the enterprise with price per seat starting at $3K/yr for Lite and $10k/yr for full platform access, to give a small handful of users a powerful amount of leverage on use cases where depth, precision, auditability are make-or-break.** "Where we compete with horizontal players is around getting work done. We go deeper, with reasoning, synthesis,

and creating content. Several of our clients use Glean for general enterprise search and use Hebbia when they need to actually do work over the information and documents they've found, running diligence, contract analysis, memo or pitchbook generation etc."

- **While consumer AI has centered around chat popularized by ChatGPT, enterprise AI has moved from chat-based assistants & copilots for search & summarization to agents which have a proactive, goal-based orientation to generate an actual outcome—whose design & orchestration gets into the nitty gritty of customer workflows and makes it resilient to competition from OpenAI itself.** "When you think about enterprise-grade AI tools, we're firm believers that enterprises will start moving away from chat. Chat just isn't the best interface for enterprise work. You don't want to iterate inside a chatbot to get your answer. You want the system you're using to understand your workflow and deliver the result in as few clicks—and ideally with as few characters—as possible. Chat doesn't service that. That's why we're moving toward an agent architecture."

- **With AI adoption in the enterprise is as much about behavioral rewiring as software provisioning, Hebbia is bundling a quasi-Accenture AI integration advice & support team of ex-bankers & lawyers who forward-deploy into customers—along with the software deployment—to help customers wrangle data integration, prompts & workflows and change management.** "AI takes operational change management. I think a lot of people, naively, believe they can roll out a chatbot and immediately drive enterprise value. Sure, you might get usage: but are those questions valuable? Are the answers valuable? How much ROI are you actually driving?"

Interview

Enterprise search has historically been a really tough category. Was it simply the advent of LLMs that helped enterprise search break out and companies like Hebbia get traction?

I think, long term, enterprise search has always been a challenging market. Historically, most tools stopped at retrieval —at simply finding what you were looking for. So when you

think about value from that perspective, the process ends before the real work even begins. These tools surface documents or information, but leave the actual reasoning to humans.

That's why it was very hard for products to break out. Large language models have started to unlock some of the downstream work—the actual reasoning and execution. At Hebbia, we began thinking about how to go one step beyond that initial layer—search or summarization—to complete the entire workflow after retrieval is done.

So we built Iterative Source Decomposition, or ISD, to overcome some of the core limitations of traditional RAG [Retrieval Augmented Generation]—like narrow context windows and chunking. ISD enables full-document reasoning. Instead of stopping at summarization, it lets enterprise search go all the way through to actual work and completion.

But in terms of trivial enterprise search—maybe you just want to find a specific document—do LLMs actually improve that search experience significantly or is it really just after the retrieval where the step-change is possible?

No, definitely on the retrieval side as well. Think about identifying the right documents, file names, and systems—it's about building agents that federate across those sources, generate the right API and search calls, perform accurate entity resolution, and map natural language inputs to document context. One area where Hebbia has gone above-and-beyond is metadata profiling. As we index over documents, we use large language models not just to process the content, but to reason over the metadata—enabling a much deeper understanding of each document. From the metadata layer you can often answer questions like: what is this document's purpose? That's we also leverage metadata to accelerate and refine search.

Turning to the customer—can you walk me through who the typical user is? Is there usually a core team or persona that's using Hebbia within an organization or is it capturing roles across functions?

We see two core wedges. I'd break customers down into two categories: the doers and the knowledge workers. The doers are operations teams. They're running work at scale, so they

want to run retrieval, search, and processing in an operational context to keep the business running. They're the ones building workflows for legal doc analysis, procurement, contract analysis, investment memo processing, building out document libraries and running extractions, synthesis, insights on top.

Then the second group is more of the knowledge worker: investment analysts, associates, MDs, lawyers, credit analysts, consultants doing diligence. These are the folks who want to go deep into their document libraries and information sets.

You've mentioned legal and investment/financial services use cases. It seems Hebbia puts a big emphasis on specific verticals? Is that intentional?

It's very intentional—at least for now. Hebbia's evolution from RAG to ISD—or Iterative Source Decomposition—let us move beyond isolated passage retrieval to robust reasoning over full documents. That was key to establishing product-market fit in high-stakes, high-complexity verticals like finance and law.

These users need accuracy. They need precision. And they need to reason across entire documents, not just search with "control-F." In a due diligence context, for example, you're not just asking "Who is the CFO?" and scanning for a name. You're trying to understand whether a company is a good investment, and whether the CFO or director affects that judgment in any way. RAG can't support that level of analysis —ISD can.

Is there anything you can share about possible market expansion—whether that's going beyond the doers and knowledge workers, or beyond finance and law use cases?

Sure, so one thing I'd emphasize is that for the time being we are really focused on finance and law, especially across financial services transactions. Your M&A and investment processes that touch financial institutions, legal institutions, and advisory institutions—your investment banks and similar. But also the consultants doing diligence or transactional advisory.

Where we're seeing interest outside of these use cases is in corporate finance functions: investor relations teams, corporate

finance teams, in-house legal at Fortune 500s. And within that same Fortune 500 market, we're very interested in R&D workflows. There we're seeing interest from companies in life sciences, oil and gas, asset maintenance—anywhere reasoning is needed over large amounts of unstructured data to get insights and drive better decisions. These are markets we'd love to pursue, but for now, we're focused on finance.

Matrix seems to be the flagship product, highly showcased on your site. Is it the flagship? And what role does it play across product and go-to-market?

Matrix really is the flagship product and in a way the engine behind the rest of the products. Post-GPT 3.5 [November 2022] we really started focusing on complex workflows. We wanted to build agent infrastructure.

And Matrix is a tabular agent interface. It's powered by a proprietary decomposition engine—that's how we go about creating the Matrix—and that enables full-document comprehension and multi-hop reasoning across different documents and data you're reviewing. So Matrix is both a default product for users who want to work in a spreadsheet-style format, and also our agent infrastructure for workflows.

That's how you start building end-to-end orchestrations for the work you do every day—like automations triggered when new documents are dropped into SharePoint or new data appears in your expert call libraries or CRM? Matrix is the agent framework for reasoning over that data and your research docs to create an output: meeting prep notes, a memo, a pitch deck from a new 10-K or earnings call, etc. Matrix becomes our agent configuration layer.

It's powerful because a lot of our users already work in tabular formats and spreadsheets. So building a tabular agent and then giving it the ability to power end-to-end workflows is really resonating.

How much lift is involved on Hebbia's end? Do you have to embed deeply with customers—do the sort of forward-deployed engineer thing? Would you go as far as saying there's a services component to the business, whether or not it's baked into contracts?

The short answer is yes. We see huge value in our engagement team. It's full of ex-professionals from the industries we serve. They're essentially designing best-in-class templates and configurations of these agents and then supporting onboarding.

When you think about enterprise-grade AI tools, we're firm believers that enterprises will start moving away from chat. Chat just isn't the best interface for enterprise work. You don't want to iterate inside a chatbot to get your answer. You want the system you're using to understand your workflow and deliver the result in as few clicks—and ideally with as few characters—as possible. Chat doesn't service that. That's why we're moving toward agent architecture.

Our engagement team, coming with their domain experience, can configure these verticalized and use-case-specific workflows so that we can land quickly. That's one piece of it.

The other piece is: AI takes operational change management. I think a lot of people, naively, believe they can roll out a chatbot and immediately drive enterprise value. Sure, you might get usage: but are those questions valuable? Are the answers valuable? How much ROI are you actually driving?

To bring a new system and a new way of working into a business, especially in financial services and law, takes real operational change. We built that capability in-house to make sure we can support clients through it. Just like you've seen with professional services firms like Accenture—they're growing AI consulting revenue by tackling this same problem. We decided to do it in-house to help customers get the most out of the product.

Maybe it's helpful if we dig more into how you define agents, both internally and for your customers. I think there's still confusion: what's an agent, what's just agentic AI, what's just an AI-automated workflow. How do you define it and how does Hebbia fit into that definition?

Great question. Everyone has their own definition. At Hebbia, the antithesis of an agent is something that's just an LLM wrapper. So anything that's a wrapper over a single large language model doing one task—we wouldn't call that an agent, even if others might.

We define agents as goal-oriented systems that can reason across multiple sources, specifically documents, and then run sub-tasks with other agents or tools to synthesize outputs or work for a user. The key characteristics are the goal orientation, the reasoning over source information, and the ability to orchestrate subtasks across functions or systems to generate an outcome.

Within Hebbia, each agent is really a coordinator of Matrix-style queries designed for specific domain objectives. For example, to run diligence over a virtual data room, where you've got contracts, models, financial statements, governance docs—it's a lot of heterogeneous info. The agent orchestrates subtasks: retrieving info, reasoning across formats, maybe running calculations, and comparing against past deals or memos. Multiple agents might handle different pieces of this, while a coordinating agent oversees the process. Then there's always with the ability for humans to interact, iterate, or redirect.

So just to paraphrase: Hebbia comes in, helps teams adapt their workflows and build around agents. You provide templates to get them started—say, agents to screen a VDR. Then it's about training the customer to coordinate and modify agents going forward. Is that right?

That's absolutely right. A lot of the templates we build are ready to go and out-of-the-box. But everyone works differently. For example, everyone looks at a credit agreement or a CIM [Confidential Information Memorandum] slightly differently—even within the same firm. Even different teams within a bank—TMT, healthcare, generalist—do the work differently. So we provide them with the right starting templates, with a lot of the prompt engineering already done. Then teams can adapt them to their workflow.

Now we're launching a new product: you drag-and-drop an output—say, an example memo or deliverable—and we'll deconstruct it and build an agent to replicate that deliverable. Then you can swap in a new company and reuse it. It empowers users to build agents based on what their MD [managing director] typically wants to see.

Here's one of my provocative questions: do agents really work? And to what extent? There's a lot of skepticism

about how accurate and reliable they are, especially as the number of steps increases and errors start to compound.
Yeah. I was about to say, the most important piece here is accuracy and precision, because error rates do compound. If the first agent picks up a 10% error, by the time you've gone through handoffs, it's like a game of telephone. You're completely off base.

So we firmly believe that agents work. But they need three things: high-quality data (so, strong retrieval), well-scoped objectives (putting the agent on rails), and strong evaluation so they can check themselves against how it's been done before. That feedback loop is critical.

And then you also need support for human-in-the-loop. How do you bring humans in when intervention or iteration is required? That's why we emphasize transparency, control, auditability, citations, data lineage, all of it. You have to build trust.

If users can't see what the agent did and why, they lose confidence. If they lose trust, usage dies. So we've built for that trust from the beginning.

Also, I'd add that we're building agents for probably one of the hardest markets out there, just given how high the accuracy bar is. In other sectors, you can tolerate a 25% error rate. We can't. So, do they work? Yes, but it depends on the problem, the use case, and how the system is built around it.

Let's talk security—and I'll lump in permissioning, privacy, and transparency. How high do these rank in terms of buyer considerations as they consider vendors?

Number one. From day one, we've been enterprise-grade here. SSO, role-based access control, audit logs, zero training on customer data. We're servicing highly regulated industries. That means meeting them where they are. Whether that's in terms of deployment, compliance, or just the expectation of baseline security and audit. It's a nonstarter otherwise.

Okay, let's move on to revenue. What can you share about growth?

I will keep it broad. Revenue and growth have been very strong and in line with what you'd expect for a best-in-class

B2B generative AI company. We're growing 100%+ year-over-year and looking to accelerate even further over the next 18 months.

How about pricing? Glean, we've heard, is roughly ~$20/seat/month. General-purpose prosumer AI tools are even cheaper. I assume you're at the higher end of the pricing spectrum?

Yeah, for sure. We have three SKUs.

There's a professional tier: that starts at $10,000 per seat, per year. That's full platform access, unlimited reasoning, workflows, automations, and all the premium integrations—PitchBook, CapIQ, broker research, etc. These are the power users who build and configure agents for the business.

But then there's the Lite product: this starts at $3,000 to $3,500 per seat, per year. These users interact with the matrix agents built by others. They also get access to our chat, deep research, and enterprise search over internal systems: SharePoint, CRMs, and so on.

So what you get is great collaboration between power-users building agents and workflows and Lite users consuming outputs and contributing to workflows.

When it comes to competition—especially against horizontal players like Glean, Microsoft Copilot, Dropbox Dash—does Hebbia's vertical strategy give you greater defensibility? Or is it just a raw product advantage that you're banking on to set you apart?

Yeah, it's a good question. Where we compete with horizontal players is around getting work done. We go deeper, with reasoning, synthesis, and creating content. Several of our clients use Glean for general enterprise search and use Hebbia when they need to actually do work over the information and documents they've found, running diligence, contract analysis, memo or pitchbook generation etc.

That's why we can be complementary, even if we see them in deals. Where we really lean in is information retrieval-plus-reasoning. We're building the infrastructure to make the context window effectively infinite. That improves accuracy and decision quality.

By the way, we don't believe in vertical-specific fine-tuned models. We're model-agnostic. The key is pairing world-class retrieval with top-tier models and then building the agent configuration and orchestration layer. That's where we differentiate.

Just to double-click on that. Say OpenAI releases a document retrieval tool for enterprise, how do you maintain an edge?

If they match us on retrieval, you still have to ask: what's the context window on their best model? Are they going to push that into the hundreds of millions of tokens? Will they invest there?

Even if they get there, you still need orchestration: agents that break problems down and manage smaller context windows. That's where Hebbia stays relevant: as a tool for models and agents, even when the models catch up. That end-to-end orchestration is where we shine.

The other piece is enterprise-grade capability: collaboration, permissioning, and full workflow support across private and public data.

How does the effectively infinite context window come about? Is it just agents stringing together multiple smaller context windows?

Basically, yes. It's what I mentioned earlier in terms of ISD or Iterative Source Decomposition, that's our proprietary architecture. It's an agent framework that enables deep document traversal and synthesis at scale. It treats question answering as a process, not a lookup.

Your CEO wrote about human-, hybrid-, and agent-first enterprises. What does that mean, and how does Hebbia power that future?

Yeah, I mentioned this earlier. Hebbia builds for hybrid workflows where humans are in control but deeply supported by AI that reasons, synthesizes, and scales across processes. We're augmenting judgment, not replacing it.

But we are starting to see a shift. Repeatable tasks are moving toward agent-first execution: parsing credit agreements, building diligence matrices, scanning filings overnight and synthesizing their impact on a portfolio. They're doing all this execution of work, not just summarizing documents.

Our long-term goal is to make Hebbia a daily tool for knowledge workers. Like Excel: trusted and indispensable. And we think we'll start seeing teams track human-to-agent ratios. Do we have enough people managing the agents running our core workflows?

And who should manage those agents? Not just tech, but domain experts, like investors, who understand the process. We're seeing that shift happen now.

Last question: what percentage of work is actually done through the Matrix or tabular interface versus other interfaces?

So all work actually passes through that Matrix-type architecture. That said, not all end users will have to, or want to, interact with that tabular view.

If you're working in a tabular process and want to work in that spreadsheet format because that's how you think or that's what fits the workflow, then you can use Matrix as your workbook once the information retrieval is done and consolidated into that view.

Alternatively, you can just interact directly with the output artifact using natural language and prompts. But behind the scenes, even if you're using another interface, the system is still running those documents, often thousands, through the Matrix architecture and doing the heavy lifting in table form.

Disclaimers