Re(AI)magine Conversations

Persistent Podcast | Episode 20

BFSI's Real Moat is Context, Not the Model.

Barath Narayanan and Ray Wang

Barath Narayanan: Hello, everyone. Welcome to Re(AI)magine Conversations, a conversation that is honest, grounded and focused on technology innovation evolving across various enterprises. These conversations go beyond the decks, beyond the buzzwords, beyond the hype. I'm Barath Narayanan:. I'm responsible for BFSI business globally at Persistent.

I spend most of my time connecting with customers across capital markets, global banks, asset management firms and payment firms — firms that are investing significant time moving from AI pilots to production. The focus now is on scaling AI adoption across the enterprise, but more importantly, doing it coherently.

The focus today is this word: coherent. Joining me today is Ray Wang, one of the leading voices on enterprise technology, AI and business model innovation. Ray, great to have you on this podcast. You're not new to this industry at all, but I'd still like you to briefly introduce yourself.

Ray Wang: Barath, thank you. Thanks for having me on this podcast. I'm Ray Wang from Constellation Research. We're a Silicon Valley-based research and analyst firm. We've been spending a lot of time looking at where AI is headed, how it's changing business models, and more importantly, how people are getting real value out of AI.

Barath Narayanan: Thank you, Ray.

BFSI clients have moved faster on AI than many other industries — that's our point of view. Whether it's agents in production, the biggest banks are committing almost $10 billion a year to AI. When you sit with BFSI leaders today, what strategic shifts are you seeing across the clientele you speak with?

Ray Wang: I think the big news is the realization that this is a business model shift. This isn't about running an experiment or a POC anymore. It's real this time — people have to get something out of this.

The second piece is that this is about decision velocity, not just technology. How can we make faster decisions and achieve better results? That could mean assessing counterparty risk faster, understanding where hedging should occur, confirming we're meeting a regulatory framework or even executing a trade faster or identifying a new market segment before someone else does.

There's so much cycle compression happening right now, and that compression matters. A good friend of mine always says he can beat any grandmaster in chess — all he needs is two moves for every one. What AI gives you is ten moves for every one.

It's also a scaling game. We're not talking about a 10% or 30% improvement. The leaders in BFSI are getting 10X, 100X, in some cases 1,000X — and you'll never hear those stories from them, because they've already achieved those results.

Barath Narayanan: That makes absolute sense to me. I'm not sure we're seeing 100X or 1,000X yet, but I can relate to what you're saying.

A year ago, it was all about which use cases to select. Today it's about running multiple use cases and figuring out how to connect the dots between them. But I agree, we're moving toward four to ten times the productivity.

Ray Wang: And the efficiency.

This is so different from digital, Barath. With digital, we were happy just to say, "it works." Now the scale is like nothing we've seen before.

Barath Narayanan: True, true. You've called 2026 the year that moves AI from promise to execution. What does this execution phase look like on the ground, especially for BFSI clients?

Ray Wang: Barath, it's the shift from the promise of POCs to the realization of results. You can keep doing POCs and that's not going to get you anywhere. The companies that have built a good data strategy, that have put the foundation and mindset in place inside the culture to get something done, are the ones seeing real results.

Why does this matter? Because we don't do AI for AI's sake. We do it to capture gains — operational efficiency, regulatory compliance, cost reduction or the execution and cycle-time compression we talked about earlier. Your ability to assess credit risk in milliseconds instead of days — that's the realization of results. Your ability to capture 10 to 20 more basis points instead of 100 basis points — that's what we're seeing across every one of our clients.

But let me ask you something. With AI in the picture, expertise is becoming a commodity. The real work is connecting the dots, understanding the business logic and handling the last mile. With expertise as a commodity, what we're seeing is that experience is not. Are you seeing the same thing — that people are looking for experience, not just expertise? The domain knowledge, the industry understanding, the people who've learned from mistakes as opposed to those who've only learned from success?

Barath Narayanan: Absolutely. There are models leading the game in how they're reinventing themselves, but everyone is catching up. All the frontier models are getting commoditized. The differentiation, as you said, is in how you connect the dots.

Most enterprises have over-invested in models and agents but under-invested in what we call "connecting the dots" — the real advantage that comes from building a semantic layer, a context and a knowledge graph. What we've concluded, Ray, is that for an enterprise to adopt AI holistically, it needs an enterprise-wide architecture.

That's where we've developed what we call the three-C framework: core, context and coordination.

Core is a secure, governing AI operating system that connects your data and workflows with full observability and strict cost control. That cost control matters, given what we're seeing across multiple customers. This is the foundation — where you manage model routing, token economics and guardrails. It's the difference between running AI and running AI you can trust and afford. That's core.

On top of core is context. This is one of the most important pieces, and it comes up in nearly every conversation we have with customers. Context is reliable, traceable access to an institution's own data, processes, rules and history, so that every agent you build has full domain awareness — who your clients are, what you are as an enterprise. It builds the lineage and compliance that should be baked into the solution.

In a bank or insurer, your moat isn't the LLM — it's your underwriting rules, your claims history, your regulatory posture. Context engineering is how you turn that proprietary knowledge into your own golden asset.

Then comes the coordination layer — a unified execution layer where people, agents and systems collaborate in a governed, auditable, process-driven workflow. This is where people, agents and systems use the context, which in turn uses the LLMs, to drive the outcomes you want.

The firms getting it right build these layers at the micro-domain level, not the overall enterprise level. If you're building context, it has to start at the micro-domain. That's the key discussion we're seeing with most clients — some are excited by it, others have already thought it through. One bank I recently spoke with has fully thought through what they call a semantic layer; some call it "the brain." But I'd say only about 3 to 4% of institutions have really thought this through so far. That's what I'm seeing, Ray.

Ray Wang: That's definitely something we're going to see more of, and I agree with you. Context is essential. Without context, how do you build trust? How do you build understanding? How do you deliver the personalization people are looking for? You can't have these systems completely isolated from each other and still expect the right conversations and the right level of understanding.

We're realizing that understanding is what builds trust and that's what allows you to work with these systems. It also protects you within the regulatory framework, which matters just as much. So, the real question is: what does a strong context layer look like in a bank or an insurer? What are the best institutions investing in for this kind of context? And looking ahead, what should leadership teams be doing now to prepare for the next phase of enterprise AI transformation?

Barath Narayanan: One failure mode we see is that when multiple agents are created and operated without a shared context, the same task produces different, inconsistent results across agents.

Ray Wang: Right — the problem is that results need to be repeatable. When you don't have consistent, repeatable results, there's a lack of trust. That's why a lot of organizations are being careful before rolling out AI — they need to be able to defend those decisions, to go back and audit that the decision trees were built properly, that the variables and attributes were weighted appropriately.

That matters a lot. If you're making a credit decision and it's not consistent every time, you could be accused of unwanted or even directed bias, because the process wasn't careful enough. These are important things to watch, especially in a highly regulated space like BFSI.

Barath Narayanan: Exactly the point. That's where a common context across agents — letting them learn from and connect with each other — becomes essential to developing a coherent outcome.

Let me give a few examples of the strong context layers we're seeing in banks and insurers.

We built a multi-agent claims flow for an insurer, covering everything from first notice of loss through fraud detection and estimation, all within one context layer. Because every agent shared a single context layer, cycle time dropped from five to seven days to under ten minutes. It worked so well because every agent references a shared digital twin of the adjudicator, so decisions stay consistent across the chain.

Another example, on the banking side: a customer doing significant modernization of their legacy payments and broker-dealer estate. When we discussed the context layer with them, it made such a difference that they brought us on board to build a context layer for this micro-domain — the payments side — within three to six months. They're seeing significant value: the effort involved is about one-fourth of what was originally estimated. That's the benefit of a context layer.

As for what leadership should do now — going back to your earlier question — don't start with hundreds of use cases. Take one use case that's holistic, end-to-end, and build the three layers we discussed. Most institutions have already invested in the core — 70 to 80% of it — and we help them close the remaining 20%. But build context before you build agents.

Ray Wang: I'd say one of the most important things is mindset. If you're using automation and AI to cut costs, you'll operate very differently than if you're trying to leap ahead. If you're trying to leap ahead, you realize this is about cycle-time compression — which, paradoxically, means you need more people, because people are now in the critical path of these decisions.

You used to send someone out for two weeks and they'd come back with a result. Now that comes back in seconds, so you need to be able to make decisions quickly. It turns out that we are the bottleneck.

So, the companies that win understand this is a business model shift — from platform to bespoke. They understand that if you've built a BPO model, it's going to be displaced. They understand this is about outcomes-based pricing — moving from insurance to guaranteed results. We're also seeing companies pursue joint ventures to achieve the same outcomes. The way you operate is going to

be very different. This isn't a normal consulting engagement or a normal services engagement — you're building new business models and partnering with services firms to succeed.

Barath Narayanan: Absolutely. Customer expectations have shifted away from "I need this specific talent" toward "here's a complicated problem I haven't been able to solve" — for example, modernizing an entire legacy underwriting or policy administration platform.

Ray Wang: Right.

Barath Narayanan: The customer's expectation now is that you understand their legacy estate, apply an AI-led solution and deliver an end-to-end value proposition. We're seeing more and more of this — in just the last three or four months, I've seen eight to ten transactions along these lines. It's no longer about individual talent; it's about outcomes.

But more importantly, the customers who treat context as infrastructure, not an afterthought, are seeing a huge differentiation. The shift is from running AI at scale to running it coherently at scale. Scale is a volume problem; coherence is an architecture problem — which is exactly what you said. The winners are the ones whose agents share context and coordinate like one enterprise. That's the real moat.

To summarize: 2026 is an execution year, where model expertise should be treated as a commodity, and the real differentiation comes from building a context layer and orchestrating coherent agents, people and systems. Ray, thank you so much for joining us today.

Ray Wang: Thank you for having me.

Barath Narayanan: And to everyone listening, thanks for tuning in to **Re(AI)magine Conversations**. If this sparked new thinking for you, follow us and let us know the guests, ideas and topics you'd like us to cover, and we'll be glad to do that.

Thank you so much.

Re(AI)magining™ the World

About Persistent

Persistent Systems (BSE: 533179 and NSE: PERSISTENT) is a global services and solutions company delivering AI-led, platform-driven Digital Engineering and Enterprise Modernization to businesses across industries. With over 27,500 employees located in 21 countries, the Company is committed to innovation and client success. Persistent offers a comprehensive suite of services, including software engineering, product development, data and analytics, CX transformation, cloud computing and intelligent automation. The Company is part of the MSCI India Index and is included in key indices of the National Stock Exchange of India, including the Nifty Midcap 50, Nifty IT and Nifty MidCap Liquid 15, as well as several on the BSE such as the S&P BSE 100 and S&P BSE SENSEX Next 50. Persistent is also a constituent of the Dow Jones Best-in-Class World Index. The Company has achieved carbon neutrality, reinforcing its commitment to sustainability and responsible business practices. Persistent has also been named one of America's Greatest Workplaces for Inclusion & Diversity 2025 by Newsweek and Plant A Insights Group. As a participant of the United Nations Global Compact, the Company is committed to aligning strategies and operations with universal principles on human rights, labor, environment and anti-corruption, as well as take actions that advance societal goals. With 468% growth in brand value since 2020, Persistent is the fastest-growing IT services brand in 'Brand Finance India 100' 2025 Report.

USA

Persistent Systems, Inc.
1731 Technology Drive
Suite 700, San Jose
CA 95110
Tel: +1 (650) 481 9180
Email: info@persistent.com

India

Persistent Systems Limited
Bhageerath, 402
Senapati Bapat Road
Pune 411016
Tel: +91 (20) 6703 0000
Fax: +91 (20) 6703 6003

Persistent
Re(AI)magining™ the World

www.persistent.com