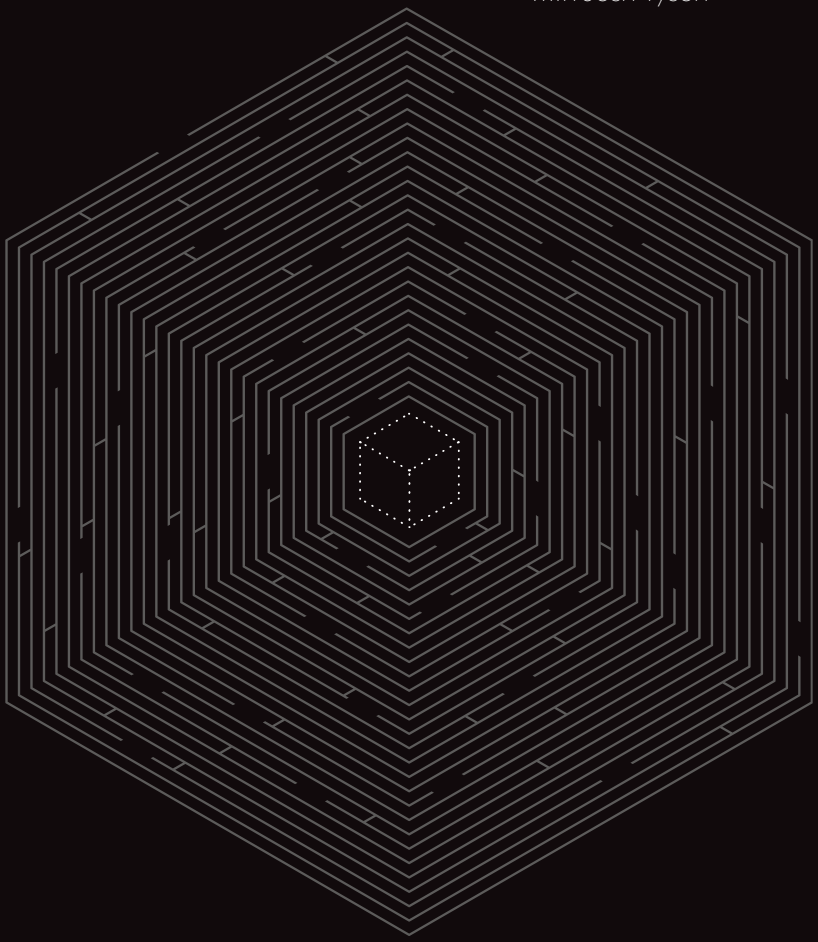
Generative AI, Point Solutions, Platforms, and Hyperglue

Fresh ideas from the authors of ***Age of Invisible Machines***, the first bestselling book on conversational AI

Robb Wilson
with Josh Tyson



INTRO

The first edition of *Age of Invisible Machines* was published a little over a year ago by Wiley. It landed on shelves two months before OpenAI released ChatGPT and introduced the world at large to conversational AI. This was essentially the adoption moment we predicted in the pages of our book, which quickly became a *Wall Street Journal* bestseller and a trusted field guide for leaders and practitioners in the AI marketplace.

The opportunity to craft a revised second edition has presented a very welcome challenge. The core information in *Age of Invisible Machines* remains solid, but the marketplace has changed in the wake of public awareness of, and interest in, generative AI. The second edition of our book will provide an updated overview of this disrupted space.

This sudden adoption of conversational AI has led to a glut of point solutions in the marketplace and we wanted to share some of the ideas we've been discussing in preparing the second edition. These generative tools are extremely powerful, but as bolt-ons they only represent a small fraction of what this technology is capable of. The way forward is to provide generative AI deeper context and orchestrate it in ways that can complete real work. This is hard but necessary work for any organization that wants to cross over into the new world that's right in front of us.

We hope you enjoy this advance look at some of the thinking that's going into the revised second edition of *Age of Invisible Machines*, available from Wiley in 2024.

— Robb Wilson and Josh Tyson

Follow this link to stay updated on the Second Edition:

Generative AI, Point Solutions, Platforms, and Hyperglue

Generative AI led by ChatGPT stirred the general public's interest in the technologies surrounding hyperautomation, sending the race toward adoption into overdrive. This has added to a glut of point solutions in the marketplace that promise to deliver on the power of these technologies. There are GPT-enabled tools popping up like crabgrass. There are also more sophisticated toolsets. Google Workspace is loaded with generative AI to help users write emails, summarize content, and create presentations (which it can also generate imagery and video for). These capabilities edge closer to using AI as more of a full-fledged solution, but there's still a ton of ground to cover.

The main deficiency is in the ability to complete real tasks. Google Workspace seems poised to cut a lot of tedium from typical productivity models, but how much work can it do outside of its own box? Next-level applications of AI employ code-free creation tools to bring outside technologies and data sources into automated workflows that use GPT (or other NLP/NLU) as a conversational interface.

Large language models like GPT can explain complex things (like what snippets of code are meant to do) but they have a very limited sense of time and place and lack the agency to do stuff. I can ask ChatGPT to draft an article for me, but I can't ask it to publish the article on my blog at 6am PST and send a summary of the article in an email to my subscribers an hour later.

As I've hopefully made clear by now, conversational AI is incredibly complex and requires an organization-wide standard of communication. It also requires the orchestration of the market-best technologies of any given moment, which requires an open and flexible platform. Hordes of point solutions won't manage the lift.

If you want to use generative AI to hyperautomate you need to be able to complete real tasks. No matter how stunning generative AI might seem, your current technology environment isn't suited to maximize its true potential. You'll need to find or build these critical components.

Critical Components of Platforms for Hyperautomation

A Contextual Memory System

The next big leap with generative AI won't be improvements in predictive power, it will come with context. This context can't come from NLU tools alone. A contextual memory system collects data from every conversation within an organization, across all channels, leveraging structured and unstructured data. This makes it possible to create channel- and user-specific experiences. Within this system, LLMs allow for rapid analysis of unstructured data, like emails, text messages, and recorded conversations. GraphDB or relational databases establish the relationships between data points, utilizing context that may not be assigned to the specific user, but is found in related datasets.

This information can be captured automatically in the form of a biosketch. Each user's biosketch can be updated in real time as they interact with the system more and more. (Users should also have control and input over what appears in their biosketches.) With hyperautomation, the perfect conversation is one where the machine correctly anticipates what the user is going to do next. It can do this by accessing the wealth of context provided by a cognitive engine. In this new paradigm, asking them, "How can I help?" should be a fallback position.

A Cognitive Orchestration Engine

Not all cognitive services are equal, and the pace of change is so fast that placing a bet on a single vendor guarantees suboptimal performance. I've also heard this referred to as cognitive

architecture, but a cognitive orchestration engine can design experiences using both legacy systems and new market-best solutions. To create real, high-functioning automations, it's critical that you can amalgamate language services (e.g. NLU, TTS, ASR, and localization) with other cognitive services, like computer vision and generative AI. This allows organizations to add vendors, manage cognitive services, and use them in different combinations, all in one place.

In essence, you're building a body for a generative AI brain. The many components that become part of this architecture can be removed and replaced as better solutions come to market (including the brain). However you choose to build this architecture, this flexibility is a critical component to using AI effectively.

Intelligent Communication Fabric

It's also crucial to enable the sharing of context and session information across channels and time. Gartner's CX CORE report states, "Intelligent coordination is a form of human and technology orchestration, where customer relationship understanding and empathy principles prescribe a unique set of coordinated actions to be executed across an organization, resulting in a frictionless and relevant CX."

Gartner calls for something called The Experience Membrane, which uses customer insight to develop a set of principles governing two-way communications between customers and a company. I've identified something similar I've been calling intelligent communication fabric (IFC). Whatever you choose to call it, you need some version of this to design experiences based on an awareness of every action taken across any channel, in real time. This fabric is designed specifically for communication. It's different from a data mesh in that you're not restricted to making right turns on a grid. Lines between points can take whatever shape makes the most sense. This fabric enables composable micro services to use conversational interfaces and conversational memory across deep channel integrations to create personalized experiences that reward users in major ways.

With a contextual memory system, a cognitive orchestration engine, and intelligent communication fabric, you can pick your preferred acronym: GPT, LAMDA, another LLM, a different NLP/NLU system? All of them are at your disposal. Same goes for existing software that's part of a manual workflow. Using APIs, they can be baked into the skills IDWs use inside your ecosystem.

It's a Bit Like Glue

My colleague, Kevin Fredrick, likes to describe these elements as "glue." At first this seemed imprecise to me. With hyperautomation, there's a whole lot more than binding going on. But then I thought of the way glue creates a teeming network of strands when you let it set between your fingertips and then slowly pull them apart.

I think I also bristled at "glue" because it suggests that everything is broken. But then it dawned on me that more things are broken rather than not inside most organizations. Technology isn't working in ways that are even remotely close to what hyperautomation requires. So maybe glue is the first step. If organizations can use ICF to standardize communications, a cognitive orchestration engine to enable legacy systems to collaborate with new technologies, and a contextual memory system to contextualize automated experiences, they can begin to repair.

So I guess we're talking about an enriched super adhesive with the power to hold and transmit. Like a sort of super glue, but more super — a hyperglue if you will.

Hyperglue has to enable these capabilities

Your architecture must be low-latency.

To support a conversational interface like GPT, latency is an essential component. The response time isn't contingent on how quickly an LLM can produce a reply. For productivity use cases, ETL (extract, transform, and load) needs to run very quickly.

Architecture is deeply integrated across channels.

To use the full functionality of each channel, you need to be able to manage sessions at a macro level as well as a granular level. To achieve a true omnichannel experience, your architecture must be capable of managing multiple channels in parallel, pogo-sticking between channels, and using native components of each channel. This level of control needs to extend to inbound and outbound channels.

Integrations must be flexible at a networking level.

If your customers have a separate digital environment, you need to be able to create a fast and secure connection with low latency.

Make use of complex data.

Your architecture must process and utilize unstructured data and convert it to structured data in real time. Intelligent communication fabric lets you play with data in new and highly sophisticated ways.

High-levels of security management.

More data and more users means more security requirements.

You need complex monitoring and debugging capabilities.

Complex systems are much harder to maintain and sustain. You need a 360 view of your ecosystem so when something goes wrong, you know where to direct your attention.

You need to react to analytics in real time.

Creating automations is a highly iterative task that never really ends. Early automations often test out hypotheses about how to automate tasks. As they are activated and interacted with you'll find opportunities to improve the experiences, which can even be implemented in real-time.

Human in the Loop:

It's critical to have humans playing active roles in the development and evolution of AI-powered automations. Part of responding to analytics in real time is letting humans come into the experience when automations need a helping hand moving forward. This kind of training brings together machines and humans for co-creating, or "co-botting" as I'm fond of saying.

Multidisciplinary teams working together.

This soft element of your architecture might be the most important. The key to creating automations that benefit customers and employees is working across disciplines and departments. Find the people who understand the processes you're trying to automate and make them part of the design process. You'll be able to build out your orchestration architecture faster with a shared vision for how people and machines will work together.

Bake-Offs: The New RFP

As a valuation tool, the request for proposal, or RFP, is being jettisoned across industries in favor of bake-offs because the latter approach offers a more hands-on, efficient approach to finding solutions that will meet organizational needs. This is especially true if your goal is hyperautomating. The fastest way to find out if and how a solution can be applied to the problems you want to solve is to see it in action, solving those problems. Comparing the same proof of concept baked on multiple platforms is a great way to find the one that fits.

That's not to say that you shouldn't be thorough. There's a reason that RFPs often stretch past the 100-page mark, and hyperautomation is no less complex than other technological endeavors. (On the contrary—it's uniformly more complex.)

Ultimately, you're in the midst of a process that requires flexibility and speed. For many organizations that haven't yet fully adopted faster, more iterative models for operating, adding a bake-off to the vendor selection or procurement is incredibly valuable, but they're not able to fully replace RFPs with them.

Any platform worth its salt will be capable of propping up a sample experience surrounding your needs (especially if they are putting their product to work internally). If they can't do that, chances are their platform isn't going to lend itself well to hyperautomation. Remember, hyperautomation hinges on design input from people with varying technical abilities working across your organization. If a solution can't be activated quickly and easily and without heavy technology lifting, it likely won't work well or won't be fast (or both).

As you interact with different vendors, you should ask how much of their own business is run by the same kinds of machines they are selling you? Whether they're a platform vendor or a services vendor, it should be easy for them to show you how they've put their solution to use for themselves.

If you've already vetted and are working with a vendor, ask yourself this: How quick are my iteration cycles? If it takes more than a week to add a new skill or iterate, you probably need to begin looking elsewhere. Having the right tools in place is paramount to building a functional ecosystem for your IDWs. It's critical to have this information and perspective to avoid getting locked into long-term "solutions" that won't be able to meet your expanding needs.

Don't Overlook UX

The biggest contributor to the abandonment of hyperautomation efforts will be poor user experience—both in the customer-facing solutions and internally. A collection of point solutions is a great strategy for abandonment. With something as all-encompassing and far-reaching as hyperautomation, usability is the number-one factor affecting adoption. You're better off not attempting hyperautomation at all than going after it with the wrong tools because if people don't adopt your solution, it won't work for hyperautomation.

I love a good challenge and, many years ago, ran headlong toward one of the worst experiences people routinely have with technology: interactions with IVRs. Nobody likes dealing with them, and this has long been an underserved area (with no significant innovations since the 1970s)—perhaps because no one had a vision of how to take it on.

That is definitely starting to change. In their 2023 Hype Cycle for Generative AI, Gartner predicts that by 2026 "more than 80% of enterprises will have used generative AI APIs or models, and/or deployed GenAI-enabled applications in production environments, up from less than 5% in 2023."

This is the nature of the types of disruptive technologies that contribute to successful hyperautomation: they are growing in strength and potential at a highly accelerated rate. A system with

the user's needs as a primary concern needs to be open to outside technology because the best solutions for optimizing experiences within an ecosystem built for intelligent hyperautomation could come from anywhere.

Right now there's probably a company you've never heard of somewhere in the world designing a tool that you will need to give your users the best automated experience. With an open system, you can incorporate that tool the moment the need arises and begin iterating on how to use it most efficiently. With a closed system, as problems emerge that can't be solved with the internal tool set, you're forced to wait for your vendor to build a solution—a purgatory that can quickly derail key business initiatives.

In this sense a closed system has a very low standard for usability. In the broken chatbot landscape, sales and marketing use their budget to start conversations with customers, and call centers are hurling money at bad automated solutions in an effort to avoid conversations with customers. I designed an open platform driven by experience design thinking so that it would be easy to create scenarios where every conversation becomes an opportunity rather than a pain point.

Bottom Line

Achieving a state of hyperautomation requires some version of hyperglue. It contains the building blocks for a bottomless fleet of IDWs that can automate increasingly complex tasks and eliminate tedium from people's lives. You won't be able to move into the future without it.

PODCAST

Even as Robb and Josh were writing *Age of Invisible Machines* they knew that the conversations that began in its pages needed to continue beyond. With that in mind, they launched Invisible Machines, a podcast exploring the world of AI through conversations with the biggest thinkers and doers in the space.

Invisible Machines has quickly become the world's most popular podcast about conversational AI and has featured guests like best selling author Seth Godin, Cassie Kozyrkov (Google's first Chief Decision Scientist), Tim Wood (Amazon's Principle Designer for AI and AI Platforms), Don Norman ("the father of UX"), Ovetta Sampson (Google's Director of User Experience with Core Machine Learning), Charlene Li (bestselling author and leadership coach), Don Scheibenreif (Vice President and Distinguished Analyst, Customer Experience Research, Gartner), and publisher Tim O'Reilly.

The Invisible Machines podcast is available on all major podcast platforms.

Robb Wilson

Robb Wilson is the co-founder and CEO of OneReach.ai, a conversational AI platform that has been named best product in general intelligence by CogX. Raised under the tutelage of philosopher (and family friend) Marshall McLuhan—who predicted the internet 30 years before it became reality—Robb has spent more than two decades applying his deep understanding of user-centric design to unlocking hyperautomation. He built *UX Magazine* into the world's largest experience design publication while simultaneously creating Effective UI, a full-service UX firm that competed with IDEO and Frog Design. In addition to launching 15 startups and collecting over 130 awards across the fields of design and technology, Robb has held executive roles at several publicly traded companies and mentored colleagues who went on to leadership roles at Amazon Alexa, Google, Ogilvy, GE, Salesforce, Instagram, LinkedIn, Disney, Microsoft, Mastercard, and Boeing. Robb puts the same passion into building a surfboard and renovating his home that he instills in the start-ups he routinely bootstraps without venture or third-party capital. A trusted thought leader in the realm of conversational AI and hyperautomation, Robb has played a part in creating a wide variety of products, apps, and movies that have touched nearly every person on the planet.

Josh Tyson

Josh Tyson is an author and producer who has held leadership roles for a variety of organizations, including TEDxMileHigh and *UX Magazine*. He is Director of Creative Content at OneReach.ai and the co-host of the N9K and Invisible Machines podcasts. His writing has appeared in numerous publications over the years, including *Big Brother Skateboarding, Chicago Reader, Fast Company, The New York Times, Observer, Thrasher, and Westword.*

ABOUT ONEREACH.AI

More than 20 years ago, Robb Wilson built a conversational bot named Cybil as part of an early AI research project. To those who interacted with Cybil, it was immediately obvious that accessing technology through conversation would transform society. It was just a question of when. Even seemingly simple prototypes required heavy software engineering and high levels of AI expertise. We knew adoption would be driven by great user experiences, and we could see that the tools just weren't good enough—what's worse, they were getting in the way.

After building a successful UX research, design and technology firm that competed with the likes of IDEO and Frog Design, Robb turned his focus to enabling mass adoption of conversational AI experiences. An early prototype of our platform quickly became valuable to AI leaders and innovators at global brands, and OneReach.ai was born.

The first generation of our platform was the result of:

- 2,000,000+ hours of testing and use-data
- 30,000,000+ users
- 10,000+ conversational applications
- 500,000+ hours of development

Our third-generation platform takes our complete set of no-code/low-code building tools to new heights, allowing customers to create LLMs that they can train on their data and connect to their systems for next-level automated experiences that pave the way to hyperautomation. By reducing technology barriers and cost for rapidly developing conversational applications without limiting flexibility, OneReach.ai is helping enterprises future proof their operations and change their relationship to technology for the better.

MEET YOUR IDW

See what's possible with Generative AI and Hyperautomation

Customize your own large language model (LLM) with this experiential demo from OneReach.ai. In a matter of minutes, you can create an industry-specific LLM that can be trained conversationally.

This is the first step in creating an ecosystem where intelligent digital workers can automate increasingly sophisticated tasks, elevate individuals and entire teams, free your employees from tedium, and establish the building blocks of context-rich, next-level customer experience.