Directo

Looking after your business

GLOBAL CX FORUM 30 AÑOS | IMT

La nueva capa operativa del contact center

Agentes de IA, autonomia y transformacion estructural

Julen Garritz
Directo Chief AI Officer

“Un agente de IA que habla, que escucha, que responde,que recuerda, que entiende en que momento de la conversacion está...”

Eso ya no es un lujo. Eso es el mínimo.

Si tu "agente de IA" es un prompt pegado a una API que lee un guion y no sabe donde esta parado...
No tienes un agente.
Tienes un ChatGPT disfrazado de contact center.

Del 9 al 11 de marzo 2026

RESONANCE Visión 2030

30 años

GLOBAL CX FORUM

Del demo al productivo

El camino real. Sin glamour, con cicatrices.

Del 9 al 11 de marzo 2026

RESONANCE Visión 2050

30 años

GLOBAL CX FORUM

El valle de la muerte del agente de IA

DEMO

Funciona perfecto.
El cliente dice lo que esperabas.
El agente responde de libro.
Todos aplauden.

PRODUCCIÓN

Miles de conversaciones al dia.
Clientes que insultan, cambian de tema,
mandan audios de 3 min en spanglish,
se quedan callados.

En medio: un abismo. Ahí muere el 90% de los proyectos.

Fase 1 | Los ciclos de ajuste antes de produccion

No es solo el prompt.
Es el prompt + la memoria + las APIs + las integraciones + los guardrails.
Todo junto. Cada vez que tocas uno, se mueve otro.

Ajustas el prompt
→ la API de tu core bancario tarda 3s → el cliente colgo

Arreglas la latencia
→ recortaste la ventana de memoria → el agente perdio contexto

Amplias la memoria
→ el costo por llamada se fue al doble

Fase 2 | Arquitectura de memoria

Un agente sin memoria = un empleado con Alzheimer.

Memoria de sesion

Lo que se dijo en esta conversacion. Ventana de contexto / tokens.

Memoria de cliente

Historial completo del usuario. RAG + embeddings + base vectorial.

Memoria operativa

Patrones globales. Mejor hora, mejor script, mejor tono por segmento

Sin estas tres capas: no tienes un agente.
Tienes un chatbot con buena voz.

Fase 3 | Escala, latencia e inteligencia operativa

Lo que se rompe al escalar:

Latencia: de 200ms a 2s bajo carga
Concurrencia: colas, timeouts, drops
Costos de GPU: escalan linealmente
Consistencia: 10 conversaciones iguales, 10,000 conversaciones impredecibles

Memoria de cliente

Historial El agente genera datos que ningun humano podria generar.
Conversa → Aprende → Optimiza → Mide→ Vuelve a conversar mejor

Directo en produccion:
85M intentos de llamada/dia | 25M conectadas | 10M SMS | NOC 24/7 | ISO 27001

Fase 4 | Human in the Loop + Compliance

Human in the loop = ecosistema donde cada quien hace lo que mejor sabe hacer.

"Si alguien les esta vendiendo que la IA reemplaza al humano, corran."

El humano cambia de rol:

Agente IA: 80% de interacciones con autonomia

Humano especializado: 20% que requiere criterio, juicio, empatia real

Sentinel: el agente que vigila a los agentes

- 100% interacciones monitoreadas
- Deteccion de sentimiento
- Alertas en tiempo real
- QA automatizado
- Reportes auditables (CNBV, Condusef)

El mapa del demo al productivo

1. Ciclos de ajuste

2. Arquitectura de memoria

3. Escala + inteligencia

4. Human in the loop

Saltarte una fase es empezar de cero.

Pero... ¿Por qué hay empresas que cruzan todas las fases y aun asi sus clientes siguen odiando la experiencia?

Porque hay algo que no esta en el mapa.

Del 9 al 11 de marzo
2026

RESONANCE
Visión 2030

30 años

GLOBAL CX FORUM

Contexto

El ingrediente invisible.

El contexto no es memoria

Memoria

Lo que el agente recuerda.
Tu nombre, tu cuenta,
que llamaste ayer.

= Quien es el cliente

Contexto

La situación. El para qué.
¿De qué es esta llamada?
¿Qué quiere lograr? ¿Dónde estamos?
¿Vino a pelear o a resolver?

= Que esta pasando ahorita

El doctor que dice: "Tiene 42 anos, mide 1.78, su ultima visita fue hace 6 meses."
→ **Eso es memoria.**

El doctor que dice: "Le duele el pecho, su papa tuvo un infarto a los 50, probablemente esta asustado. Vamos a descartar lo grave primero."
→ **Eso es contexto.**

El contexto como ventaja competitiva

El contexto no es un feature tecnico.
No es una linea en un RFP.

Es la diferencia entre que tu cliente sienta
que habla con una empresa que lo conoce...
o con una que lo trata como un numero.

La voz la van a tener todos.
La velocidad la van a tener todos.
La memoria la van a tener todos.

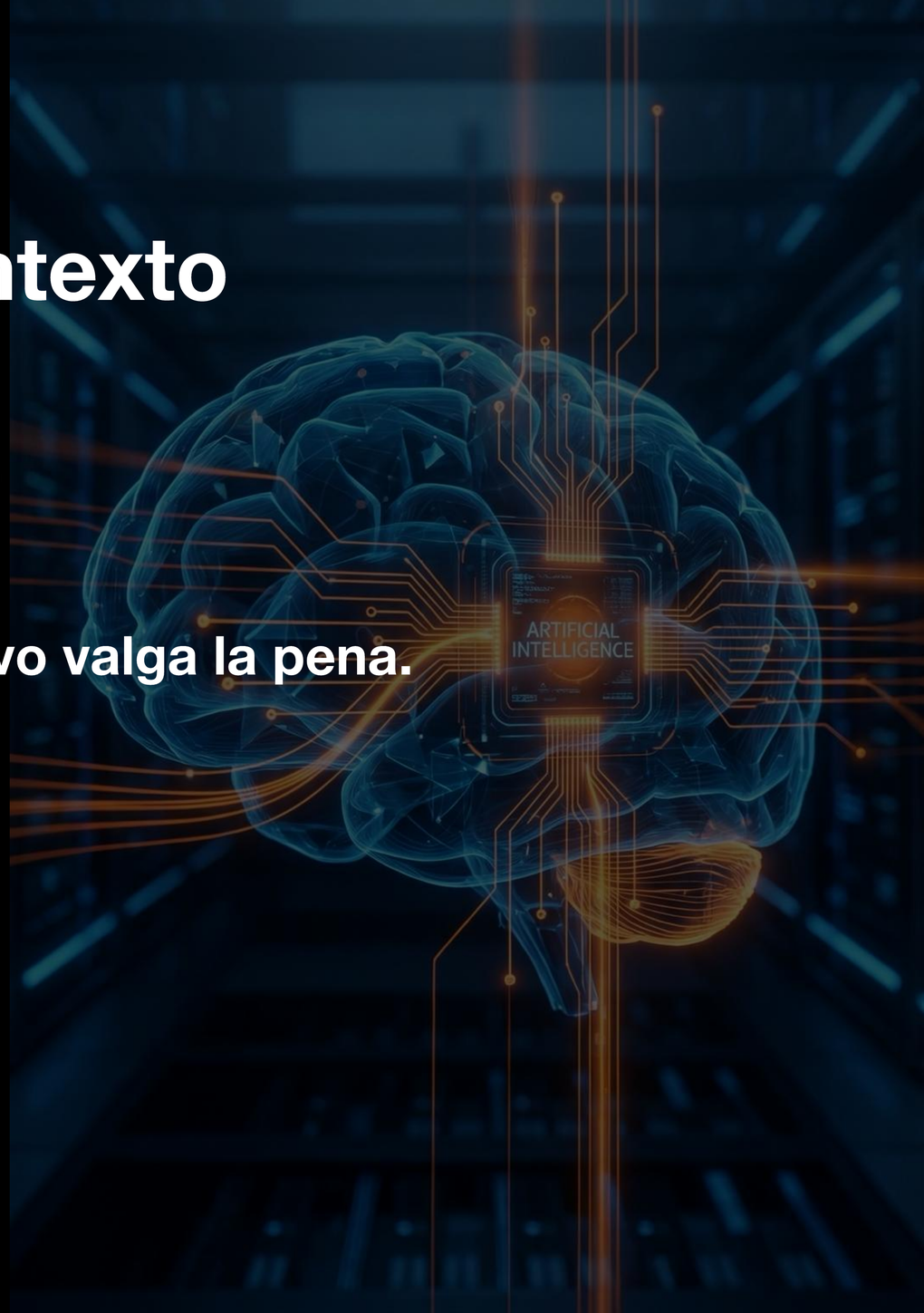
Saber en que momento estas de una relacion
con un ser humano... eso es lo dificil. Y eso es lo que vale.

Demo + Produccion + Contexto

Las fases te llevan del demo al productivo.
Ese es el como.

El contexto es lo que hace que el productivo valga la pena.
Ese es el para que.

Sin el camino, no llegas.
Sin el contexto, llegas a nada.



Del 9 al 11 de marzo
2026

RESONANCE
Visión 2030

30 años

GLOBAL CX FORUM

Full Duplex: La Conversacion Que Viene

Del 9 al 11 de marzo
2026

RESONANCE Visión 2030

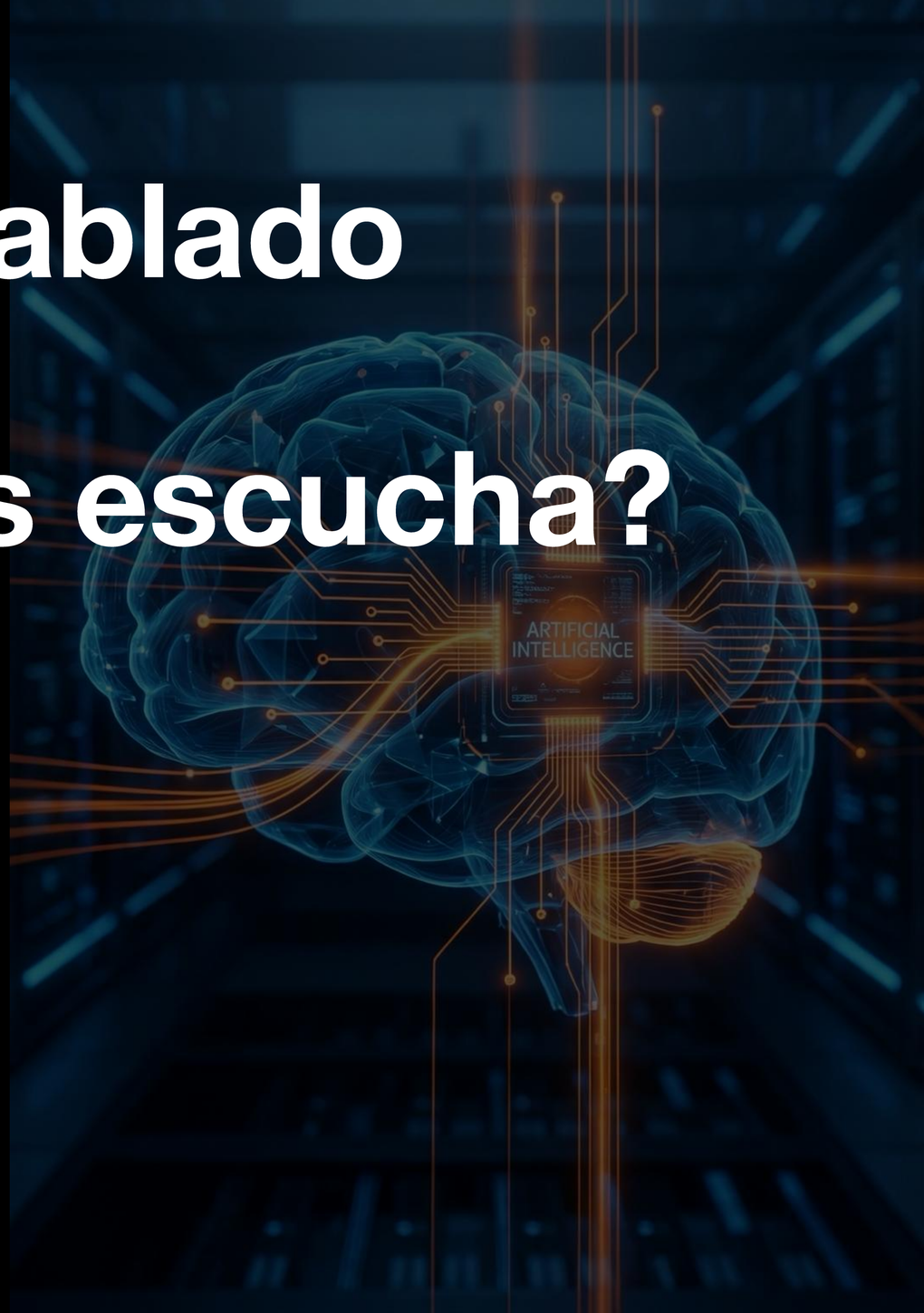
30 años

GLOBAL CX FORUM

Alguna vez han hablado con alguien que realmente los escucha?

No que espera su turno para hablar.
Que reacciona a su tono.
Que nota cuando dudan.
Que sabe cuando callarse.

Eso es lo que viene.
Y se llama **full duplex**.



Half Duplex vs Full Duplex

HALF DUPLEX
(hoy)

- Tu hablas.
- El agente espera.
- El agente procesa. Tu esperas.
- Turnos rigidos.
- Latencia perceptible (1-3s).
- No detecta tono en tiempo real.
- No puede interrumpir naturalmente.

Se siente como hablar con una maquina

FULL DUPLEX
(lo que viene)

- Ambos hablan y escuchan al mismo tiempo.
- Reacciona mientras hablas.
- Detecta emocion en tiempo real.
- Interrumpe naturalmente.
- Ajusta respuesta a medio camino.

Se siente como hablar con una persona.

Del 9 al 11 de marzo 2026

RESONANCE Visión 2030

30 años

GLOBAL CX FORUM

Full Duplex no es solo latencia. Es un cambio en la naturaleza de la conversacion.

Por primera vez, la máquina no solo procesa.
No solo responde.
No solo recuerda.

Escucha.

Detecta que subiste el tono y cambia de estrategia.
Nota que dudaste y ofrece una alternativa.
Siente el silencio... y sabe que no debe llenarlo.

Del 9 al 11 de marzo 2026

RESONANCE Visión 2030

30 años

GLOBAL CX FORUM

La nueva capa operativa: tres agentes, un ecosistema

Agente IA Conversacional

+ **80%** de interacciones con autonomia. Memoria, contexto, 24/7, omnicanal.

Agente Humano Especializado

+ **20%** que requiere criterio, empatia real, juicio experto. Ya no contesta 200 llamadas. Resuelve 40, las que importan.

Agente de Supervision IA

+ Sentinel. **100%** monitoreado. Compliance, alertas, inteligencia operativa.

El lider de CX ya no administra headcount. Orquesta un ecosistema.

Dentro de 3 años, un cliente va a llamar a su banco.
Y el agente va a decir:

"Hola Juan, se que la semana pasada tuviste un problema con tu tarjeta, que lo resolvimos por WhatsApp, y que hoy me llamas porque recibiste un cargo que no reconoces. Ya lo estoy revisando."

Y Juan va a colgar pensando:
"Que buen servicio."

Sin saber que hablo con una maquina.

Del 9 al 11 de marzo 2026

RESONANCE Visión 2030

30 años

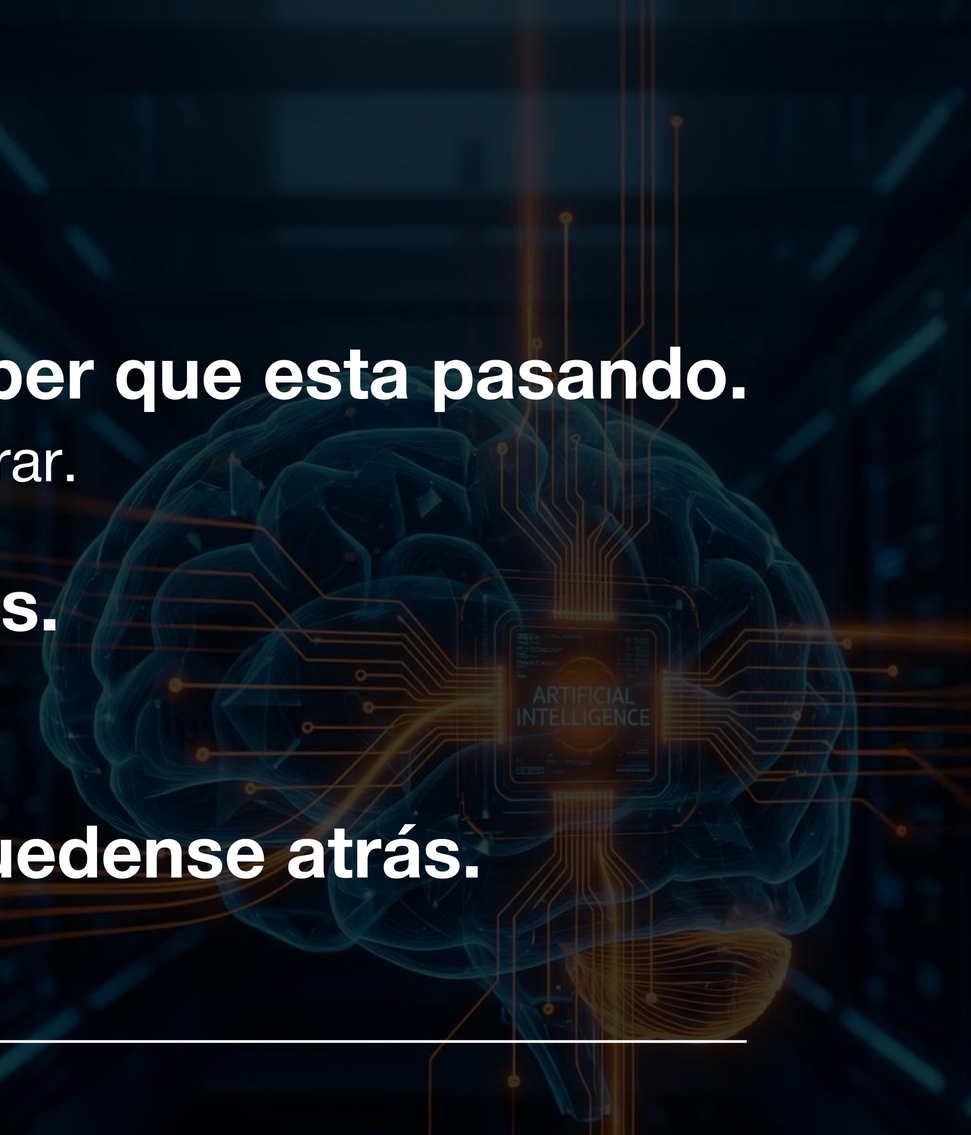
GLOBAL CX FORUM

3 ideas para llevarse hoy:

1. El contexto no es memoria. Es saber que esta pasando.
Y es la unica ventaja competitiva que va a durar.

2. Del demo al productivo hay 4 fases.
Saltarte una es empezar de cero.

3. Full duplex viene. Preparense o quedense atrás.



Julen Garritz
Directo Chief AI Officer

julengarritz.com | directo.com

GLOBAL CX FORUM

RESONANCE
Visión 2030

IMT

Directo