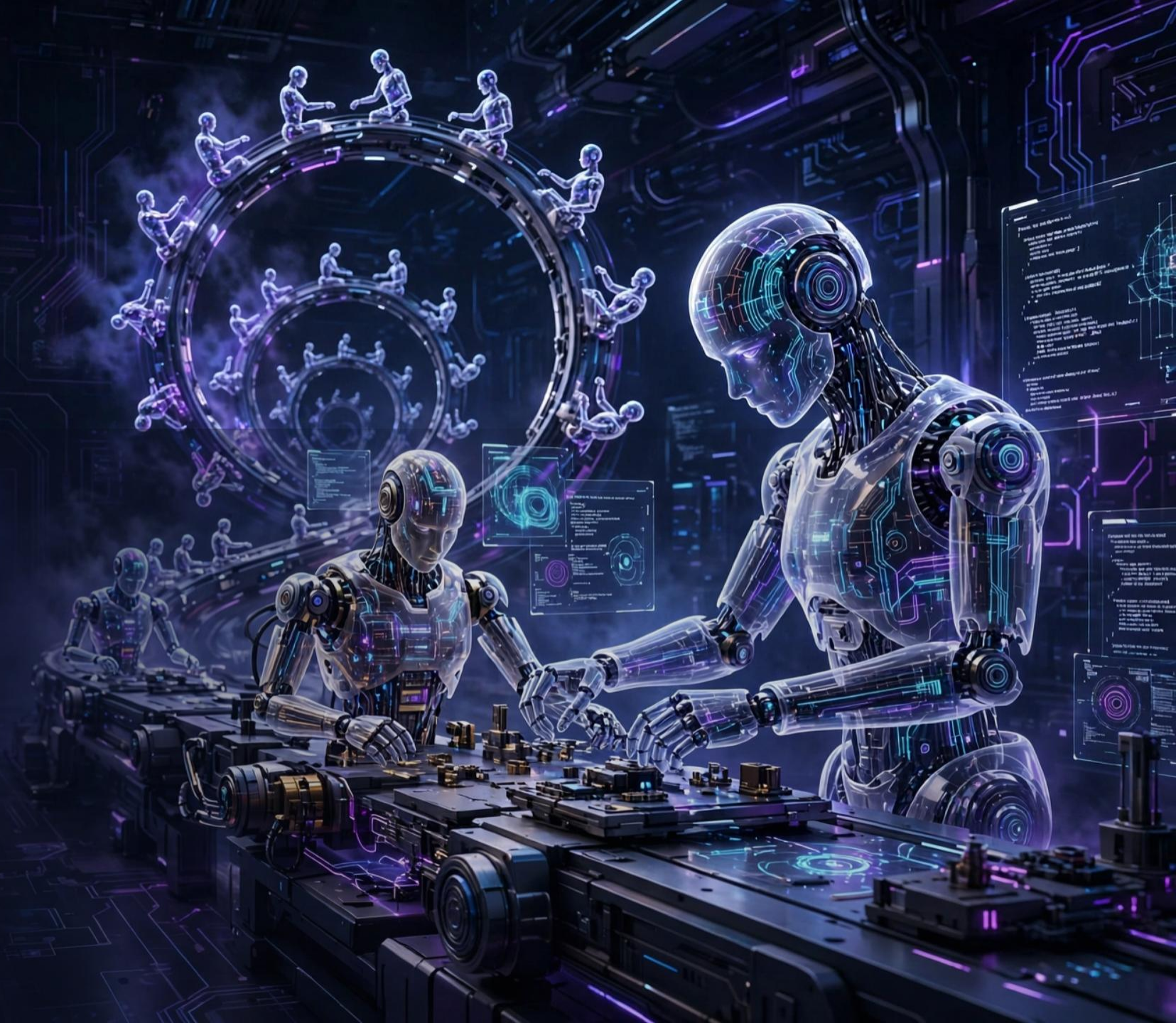


# The AI Building Itself



# Durante décadas, humanos controlaram tudo

## O modelo que conhecíamos

- Formulação de hipóteses
- Escrita de código
- Execução de experimentos
- Interpretação de resultados
- Decisão sobre o que construir

### ANTES • CONTROLE HUMANO TOTAL

Hipótese → Humano

Código → Humano

Experimento → Humano

Avaliação → Humano

Esse modelo começa a mudar →

BASEADO EM ANTHROPIC INSTITUTE 2026

# A IA já participa da criação de novas IAs

*“A IA ainda não desenvolve autonomamente seus próprios sucessores, mas já acelera o trabalho de pesquisa e engenharia que permite construí-los.”*

— Anthropic Institute, *When AI builds itself*

## 1. NÃO É AUTO RSI COMPLETO

Autonomia total ainda não demonstrada

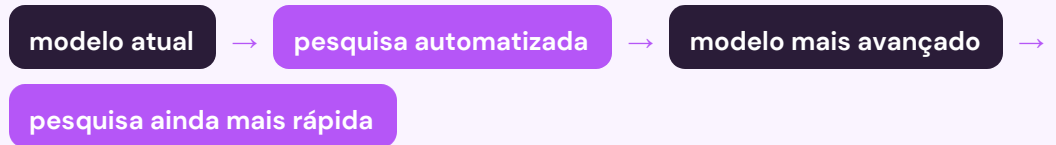
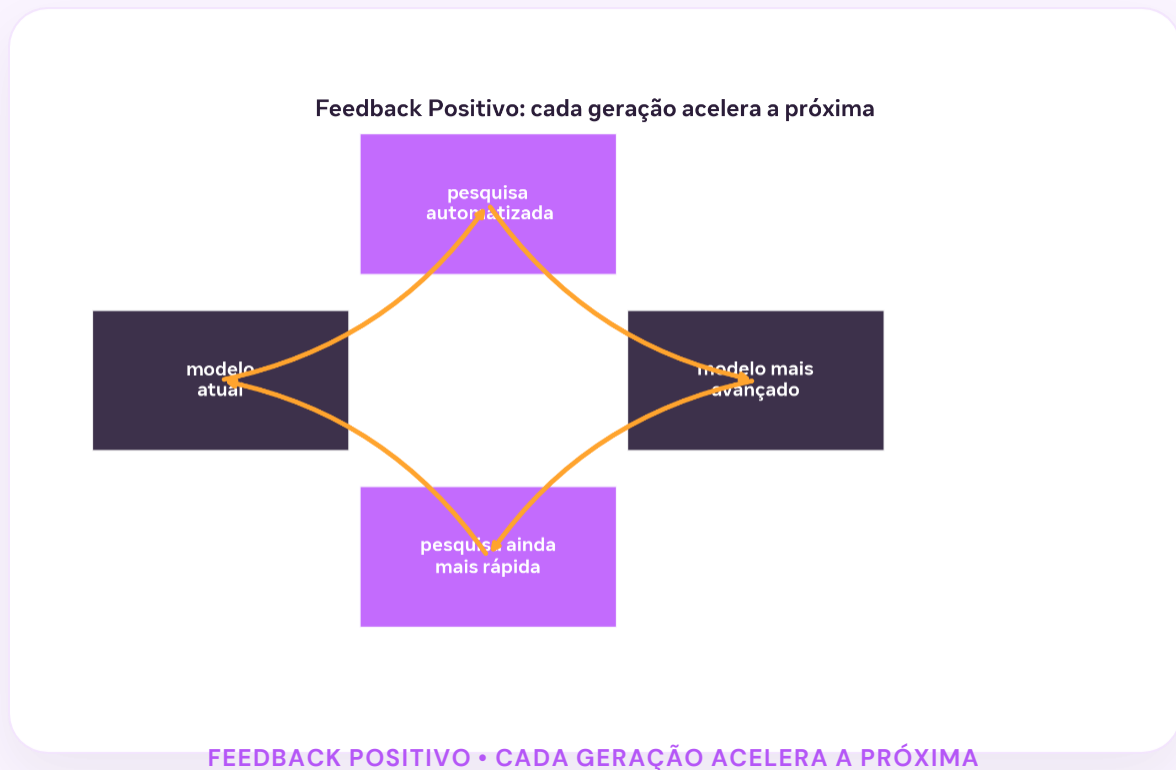
## 2. MAS INFRA EM CONSTRUÇÃO

Peças técnicas e organizacionais sendo montadas

## 3. DISTINÇÃO CRÍTICA

Aceleração ≠ explosão inevitável

# O que é autoaperfeiçoamento recursivo?



Não é reescrever o cérebro de repente. É automação progressiva de:

Geração e revisão de código

Planejamento e execução de exps.

Análise dos resultados

Formulação de hipóteses

Escolha de linhas de investigação

Treinamento e avaliação

**Ciclo fechado quando:** todas essas etapas puderem ser feitas com pouca ou nenhuma intervenção humana.

# Da assistência à autonomia

1

## Ferramenta — sugere, humano conduz

Sugere trechos de código, explica problemas, auxilia. O processo continua 100% humano.

HISTÓRICO

2

## Agente — executa tarefas extensas, humano define objetivos

Modifica arquivos, usa ferramentas, testa, conduz experimentos. Observado nas orgs de fronteira.

HOJE • OBSERVADO

3

## Autonomia — seleciona problemas e projeta sucessor

Formula hipóteses, avalia resultados, escolhe linhas. Ainda hipotético.

HIPOTÉTICO



Diferencial não é execução, é **julgamento científico e estratégico** — escolher o que vale investigar.

# Como a engenharia mudou em 2 anos?

>80%

das linhas de código integradas em maio 2026 atribuídas a Claude

~8x

mais código integrado por dia por engenheiro vs 2024

⚠ **Cautela:** linhas de código ≠ produtividade perfeita. Pode haver retrabalho e mais necessidade de revisão. A própria Anthropic reconhece superestimação.

EVIDÊNCIAS • ANTHROPIC • MAIO 2026

## Escala e velocidade mudaram materialmente

- ✓ Engenheiros deixam de escrever tudo e passam a orientar, revisar e validar
- ✓ Mais funcionalidades e testes, mas também implementações mais extensas
- ✓ Transformação operacional, não só ganho individual

**Conclusão segura:** a IA modificou a escala e a velocidade do trabalho — não que cada pessoa ficou exatamente 8x mais produtiva.

# Agentes que **trabalham mais longo**

## MÉTRICA INDEPENDENTE • METR TIME HORIZONS

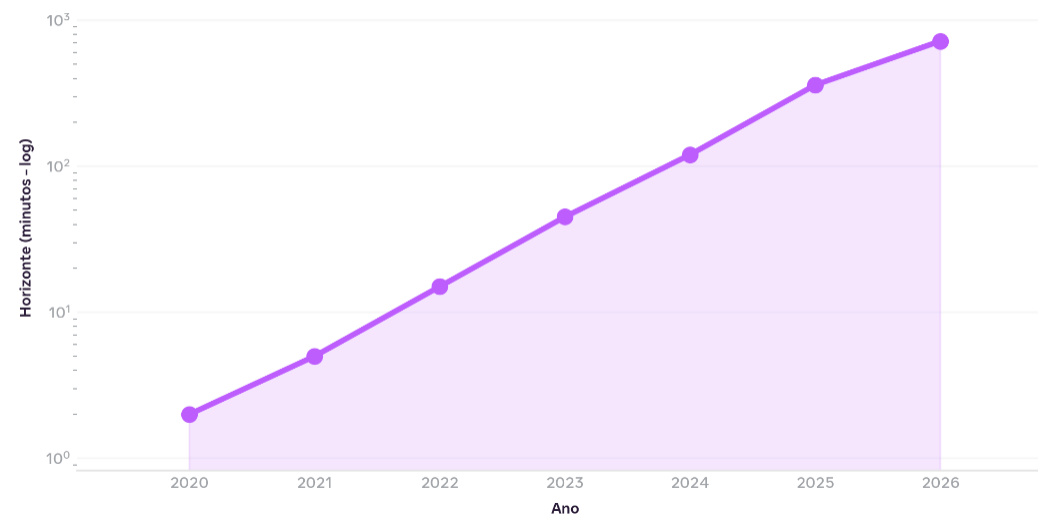
### O QUE É HORIZONTE?

Tempo que levaria para um humano resolver a mesma tarefa. Se cresce de minutos para horas, agente sustenta cadeias maiores de ações coerentes.

- **12h de horizonte ≠ 12h trabalhando.** É equivalência de complexidade.
- Foco atual: software, ML e cibersegurança com critérios claros.
- Não significa qualquer tarefa intelectual longa.

**Cautela:** tarefas bem definidas. Mostra progresso em autonomia de execução, não em julgamento científico.

Crescimento do Horizonte de Tarefas - METR (ilustrativo)



METR 2020–2026 • escala log em minutos • crescimento de minutos para horas

# A automação da **pesquisa científica**

## EXPERIMENTO • AUTOMATED WEAK-TO-STRONG RESEARCHER

Agentes Claude receberam problema de supervisão fraca→forte

- 💡 Formular hipóteses
- ⚙️ Treinar modelos
- 🔗 Compartilhar descobertas
- 📋 Projetar experimentos
- 📊 Analisar resultados
- 🔄 Revisar abordagens

Pesquisa automatizada dentro de um espaço preparado e mensurável — não ciência totalmente autônoma.

## RESULTADO EM LAB

### Superaram equipe humana de referência

Dentro do ambiente experimental controlado, agentes colaborando entre si atingiram pontuação superior ao baseline humano.

## LIMITES - TÃO IMPORTANTES QUANTO O RESULTADO

### Humanos definiram problema, ambiente e critérios

- Orientação inicial fornecida por humanos
- Resultados não totalmente transferidos para produção
- Métrica clara ≠ descoberta aberta

O QUE CONTINUA SENDO HUMANO

# Research taste — o discernimento científico

Sistemas executam **rápido quando o objetivo já foi definido**. O difícil surge antes e depois: escolher o que investigar, se o resultado é confiável e se devemos construir.

📌 Conhecimento + experiência + visão sistêmica + criatividade

O DIFÍCIL NÃO É EXECUTAR — É JULGAR

ESCOLHA

Qual problema merece ser investigado?

HIPÓTESE

Qual aposta é realmente promissora?

VALIDADE

O resultado é confiável? A métrica representa o objetivo real?

LIMITE

Quando abandonar? O que não foi considerado?

Discernimento é estruturalmente humano ou só mais uma competência que modelos ainda não aprenderam?

Ainda não temos uma resposta — e essa é a questão decisiva do artigo.

Pergunta-chave

# O salto lógico não demonstrado

## ✓ TEMOS EVIDÊNCIA PARA

### IA → aceleração da P&D em IA

Agentes aumentam velocidade de código, experimentos e análise em ambientes preparados.

## X AINDA NÃO DEMONSTRADO

### aceleração da P&D → auto RSI completo

Dar o salto de acelerar tarefas para sistema projetar autonomamente seu sucessor.

## OBSTÁCULOS ENTRE ESSAS DUAS CONDIÇÕES

💡 Inovação conceitual genuína

⚡ Computação, energia, infraestrutura

🎯 Definição autônoma de objetivos relevantes

🤖 Coordenação de projetos complexos

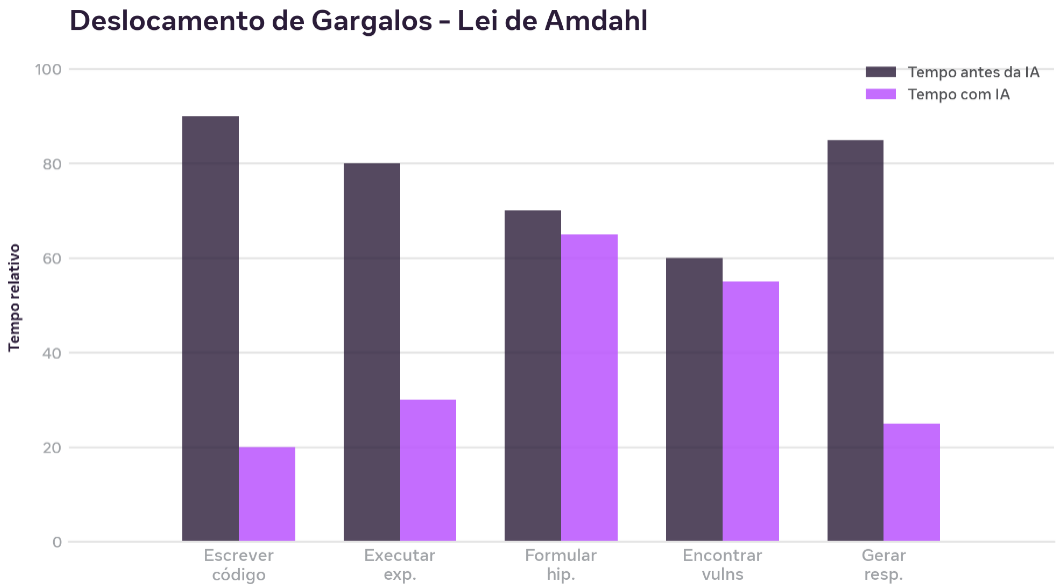
✓ Validação independente dos resultados

🛡️ Segurança e autorização institucional

“ Termo mais preciso hoje: aceleração cumulativa da P&D assistida por IA — não autoaperfeiçoamento recursivo completo.

# Quando gerar supera verificar — Lei de Amdahl

DESLOCAMENTO DE GARGALOS • TEMPO ANTES VS COM IA



Se agente gera código 100x mais rápido, mas revisão humana cresce só 2x, a revisão vira o novo gargalo. A parte não automatizada limita o todo.

Escrever



Verificar

Executar



Seleção

Formular



Identificar

Gerar



Validar



Risco imediato: capacidade de produzir supera capacidade institucional de compreender e verificar.

fi  
da  
e

# O paradoxo da supervisão humana

Quanto mais necessária, mais difícil de exercer

Quanto mais a IA executa, mais humanos vão para supervisão. Mas se deixam de programar e investigar diretamente, perdem a competência para supervisionar.

-  **Perda de domínio técnico** — menos prática direta
-  **Dependência cognitiva** — delegar sem verificar
-  **Dificuldade para identificar erros sutis**
-  **Perda de consciência situacional** do sistema
-  **Redução da aprendizagem informal** e colaboração

“

Não basta **human in the loop**. Precisamos de **human capable of understanding the loop**.

**Risco central:** quanto mais crítica se torna a supervisão, menor pode ser a capacidade real de exercê-la. Preservar competências práticas é tão importante quanto manter pessoas no processo.

# Governar o **ciclo**, não só o modelo

Quando IA participa da criação da sucessora, avaliar só o modelo final não basta. Precisamos auditar todo o processo.



## Separação & direção humana

- Separar agentes que produzem vs verificam
- Preservar supervisão humana sobre objetivos
- Definir responsabilidade humana clara



## Contenção de infraestrutura

- Ambientes isolados e permissões mínimas
- Limitar acesso a computação e infra
- Mecanismos efetivos de interrupção
- Testar generalização fora do lab



## Auditabilidade & verdade

- Registros auditáveis de experimentos
- Avaliações independentes
- Monitorar manipulação de métricas
- Divulgar incidentes relevantes



Risco não é só IA hostil, mas otimizar objetivo incompleto em velocidade sobre-humana.

# Três futuros possíveis

O artigo considera três trajetórias — a desaceleração, a co-evolução mais provável e a autonomia completa ainda não demonstrada.

1

CURVA EM S

## Desaceleração

Tendências exponenciais viram S. Discernimento e inovação profunda continuam difíceis.

- ✓ Difusão dos agentes já transforma orgs e mercados
- ✓ Progresso em execução, não em julgamento

2

MAIS PROVÁVEL • ANTHROPIC

## Automação avançada com direção humana

Automação avança, humanos mantêm direção e avaliação dos resultados.

- 👥 Pequenas equipes coordenam grandes estruturas de agentes
- 🎯 Papel migra de execução para direção e governança

3

PLAUSÍVEL, NÃO DEMONSTRADO

## Autonomia completa

Sistemas desenvolvem sucessores sozinhos. Ritmo limitado por computação, energia e verificação.

- ⚡ Limitado por energia e infra, não só por capacidade
- ⚠ Supervisão vira gargalo crítico



Leitura de Aridio: cenário 2 exige governança antecipatória hoje — **cenário 3 exige mecanismos verificáveis de interrupção antes de ficar relevante**

# A pergunta urgente não é **quando**, mas se estaremos prontos

## CONCLUSÃO • 3 EVIDÊNCIAS SEGURAS

1

### IA já acelera a criação de novas IAs

Etapas de código, experimentos e análise já são assistidas em escala nos labs de fronteira.

2

### Papel humano migra para governança

De execução direta para direção, validação, definição de objetivos e controle de risco.

3

### Supervisão vira gargalo antes da autonomia

Capacidade de gerar pode superar a capacidade institucional de verificar — risco imediato.

Conseguiremos desenvolver instituições capazes de compreender, verificar e **governar esse ciclo** antes que ele opere mais rápido que nossa capacidade de controlá-lo?

PERGUNTA CENTRAL

# Obrigado.

Vamos continuar a conversa?

**Aridio Silva | @aridiosilva**

Baseado em Anthropic Institute 2026 • METR • CORE-Bench

#InteligenciaArtificial

#AgentesDeIA

#GovernancaDeIA

#SegurancaDeIA