Oscar Obando
CHAVES

La tecnología que se construye sola.

El 4 de junio de 2026, Anthropic reveló que más del 80% de su código ya lo escribe su propia IA, y pidió al mundo la opción de frenar. Qué significa para una región que ni siquiera está en la conversación.

Oscar Obando Chaves · Senior AI Counsel · CAIO
oscarobando.ai
Junio 2026

Lo que Anthropic admitió

El 4 de junio de 2026, el Anthropic Institute publicó un documento con un título que no intenta suavizar nada, "When AI builds itself" (cuando la IA se construye a sí misma), firmado por **Marina Favaro**, directora del instituto, y **Jack Clark**, cofundador de la empresa. La tesis central cabe en una línea, la IA está asumiendo cada vez más el trabajo de desarrollar IA, y eso ya está acelerando el ritmo del avance.

El concepto que organiza el documento se llama **automejora recursiva**. Una IA diseña y entrena a la siguiente versión de sí misma, esa versión más capaz construye la siguiente, y así sucesivamente, con el ser humano corriéndose poco a poco hacia un papel de supervisión. Anthropic aclara dos veces que ese punto todavía no se alcanzó y que no es inevitable, pero advierte que podría llegar antes de que la mayoría de las instituciones esté preparada.

El concepto lleva años en la literatura técnica. Lo que convierte este documento en noticia es la propuesta de gobernanza que lo acompaña. Anthropic escribió, en una frase que conviene leer con cuidado, que sería bueno para el mundo tener la opción de frenar o pausar temporalmente el desarrollo de IA de frontera para que las estructuras sociales y la investigación en seguridad puedan ponerse al día. Es la primera vez que un laboratorio líder propone formalmente un mecanismo de freno coordinado, aunque, como veremos, con una condición que cambia todo.

Fuentes. Anthropic Institute, "When AI builds itself" (4 jun 2026, anthropic.com/institute/recursive-self-improvement). Scientific American (5 jun 2026). PYMNTS (5 jun 2026).

La evidencia desde adentro

El documento se apoya en datos internos que Anthropic nunca había publicado. El más contundente es este, a mayo de 2026 más del **80%** del código que la empresa integra a su base de producción fue escrito por Claude, su propio modelo, cuando antes del lanzamiento de Claude Code en febrero de 2025 esa cifra estaba en un dígito bajo.

El segundo dato mide la productividad por ingeniero. En el segundo trimestre de 2026, el ingeniero promedio de Anthropic integraba **ocho veces** más código por día que en 2024, porque la mayor parte ahora la escribe Claude mientras el humano dirige y revisa. La propia Anthropic advierte que esa cifra sobreestima la ganancia real de productividad, ya que las líneas de código miden cantidad y no calidad, pero el sentido de aceleración queda claro. La tasa de éxito de Claude en las tareas más abiertas, aquellas donde ni siquiera está claro cómo se ve la solución, llegó al **76%** en mayo de 2026, cincuenta puntos porcentuales más que seis meses antes.

Sobre la calidad, Anthropic es honesta. Reconoce que a fines de 2025 el código escrito por Claude era algo peor que el escrito por humanos, que hoy está a la par, y que esperan que sea estrictamente mejor dentro del año. Un ejemplo del documento da la dimensión del cambio, en abril de 2026 Claude entregó más de **800 correcciones** que redujeron una clase de errores de API por un factor de mil, y el ingeniero que lo superó estimó que un humano habría necesitado **cuatro años** para completar ese trabajo.

Fuentes. Anthropic Institute, "When AI builds itself" (4 jun 2026). Anthropic precisa que el liderazgo estimó públicamente 90% o más de código escrito por Claude; el 80% es una medición más conservadora de líneas integradas a producción.

La evidencia desde afuera

Los datos internos podrían leerse como marketing, por eso Anthropic los acompaña de mediciones públicas e independientes. La más reveladora viene de METR, una organización que mide cuánto tiempo de trabajo humano puede completar un modelo de forma autónoma. La duración de las tareas que la IA resuelve con fiabilidad se está **duplicando cada cuatro meses**, frente a una tendencia previa de cada siete.

En marzo de 2024, Claude Opus 3 completaba tareas de software que a un humano le tomaban unos cuatro minutos. Un año después, Claude Sonnet 3.7 manejaba tareas de hora y media. Un año más tarde, Claude Opus 4.6 resolvía tareas de **doce horas**. El modelo interno que Anthropic llama Mythos Preview puede trabajar al menos **dieciséis horas** seguidas. En el banco de pruebas SWE-bench, que entrega a un modelo un caso real de ingeniería de software con su reporte de error, el desempeño pasó de un dígito bajo a fines de 2023 a rozar el techo del 94% hoy.

El dato que más debería interesar a quien piensa en seguridad nacional o corporativa viene del Project Glasswing. Anthropic afirma que, en sus primeras semanas, Mythos Preview encontró más de **diez mil vulnerabilidades** de severidad alta y crítica en algunos de los sistemas más importantes del mundo, al punto de que el cuello de botella en ciberdefensa se desplazó de encontrar fallas a parchearlas a tiempo. La misma capacidad que protege puede atacar.

Fuentes. Anthropic Institute, "When AI builds itself" (4 jun 2026), citando a METR (metr.org), SWE-bench (swebench.com) y datos del Project Glasswing.

Pedir el freno mientras se acelera

El llamado a frenar llegó con un detalle de calendario que conviene no pasar por alto. Apenas tres días antes, el 1 de junio de 2026, Anthropic presentó de forma confidencial su solicitud de salida a bolsa ante la SEC, después de una ronda que la valoró cerca de **USD 965,000 millones**, con reportes que ubican la oferta pública por encima del billón de dólares. Que un laboratorio camino a esa valoración pida pisar el freno generó escepticismo inmediato.

La condición que acompaña la propuesta explica buena parte de la tensión. Anthropic condiciona su propio freno a una sola cosa. Reduciría el ritmo o pausaría únicamente si otros laboratorios de frontera lo hicieran también, y solo bajo condiciones verificables. El problema es que verificar una pausa en IA es endiabladamente difícil. Como reconoce el propio documento, los entrenamientos son mucho más fáciles de ocultar que un silo de misiles, sus insumos son de propósito general, y el incentivo a hacer trampa en silencio es enorme, porque quien siga avanzando mientras los demás se detienen hereda la delantera.

Las reacciones se dividieron. El crítico Gary Marcus calificó el anuncio como un cambio de carnada, retórica de seguridad sincronizada con la salida a bolsa. Otros respondieron que publicar datos internos sobre la aceleración del desarrollo es en sí mismo una contribución al debate de gobernanza. Desde OpenAI, Sam Altman ha sostenido que deben ser los gobiernos democráticos, y no las empresas privadas actuando solas, quienes definan las reglas. Ningún otro laboratorio se ha comprometido a sumarse al freno propuesto.

Fuentes. Let's Data Science (5 jun 2026, sobre el S-1 del 1 jun y la entrevista de Jack Clark en BBC Newsnight). PYMNTS (5 jun 2026). Gary Marcus, Substack (5 jun 2026). Sam Altman, India AI Impact Summit 2026.

Los tres futuros que Anthropic dibuja

Escenario	Qué ocurre	Implicación
1. La curva se aplana	La tendencia exponencial resulta ser una curva en S; los retornos de escalar el cómputo disminuyen. Pero las capacidades actuales se difunden a toda la economía	Una empresa de 100 personas hace el trabajo de 1,000. Es el que da más tiempo de adaptación. Anthropic lo considera poco probable
2. Ganancias que se acumulan	El desarrollo se automatiza en gran parte, pero los humanos siguen fijando la dirección y juzgando resultados	Una empresa de 100 personas hace el trabajo de 10,000 a 100,000. Anthropic dice que es el escenario hacia el que probablemente vamos
3. Automejora recursiva plena	La IA diseña y construye a sus sucesores; el humano queda en validación y verificación de un laboratorio virtual	El ritmo lo marca solo el cómputo disponible. El futuro sobre el que Anthropic dice tener menos certeza, incluido el problema de alineación

Anthropic apuesta por el segundo escenario, y ahí aparece el límite que ellos mismos sí reconocen. Acelerar una parte del proceso suele mover el cuello de botella a otra, un principio que en computación se conoce como la ley de Amdahl. La empresa ya lo experimentó, a medida que Claude produce más código, la **revisión humana** se convirtió en el nuevo embudo. Cuando el hacer cuesta casi cero, lo escaso pasa a ser el criterio para decidir qué vale la pena hacer y qué resultados merecen confianza.

Fuentes. Anthropic Institute, "When AI builds itself" (4 jun 2026), sección "Possible futures".

¿Y América Latina?

Hay un detalle que ninguna cobertura de esta noticia menciona, y es el más relevante para nosotros. Toda la conversación sobre frenar, verificar y coordinar ocurre entre laboratorios de Estados Unidos, el Reino Unido y China. América Latina no está en la sala. América Latina llega a esa conversación como espectador que consume la tecnología, sin participar en construirla, sin autoridad para regularla y sin medios para verificar una pausa.

Eso vuelve incompleto un principio que aparece en casi todos los proyectos de ley de la región y en la propia estrategia ecuatoriana de IA, el de la **supervisión humana**. El documento de Anthropic muestra que dentro de la empresa que crea estos sistemas la revisión humana ya se convirtió en el cuello de botella, y que sus propios ingenieros se están corriendo hacia un papel de supervisores de un trabajo que la IA hace más rápido de lo que ellos pueden revisar. Si quienes construyen el sistema enfrentan ese límite, conviene preguntarse qué significa exigir supervisión humana como requisito legal en una norma ecuatoriana, sin capacidad técnica para definir en qué consiste ni para auditar si se cumple.

La verificación agranda la brecha todavía más. Una pausa creíble exige poder comprobar que los demás se detuvieron, y ninguna autoridad latinoamericana tiene hoy la capacidad de verificar absolutamente nada sobre un modelo de frontera. Copiar el vocabulario regulatorio europeo, con sus categorías de riesgo y sus obligaciones de trazabilidad, sin la infraestructura para operacionalizarlo, produce normas que se leen bien y no protegen a nadie. La ventana útil para la región está en otra parte. Construir la capacidad institucional y técnica para entender, auditar y negociar con quienes fabrican la IA de frontera importa mucho más que redactar normas sobre una tecnología que no producimos, y conviene hacerlo antes de que el ritmo del que habla Anthropic vuelva esa ventana todavía más estrecha.

Fuentes. Anthropic Institute, "When AI builds itself" (4 jun 2026). Índice Latinoamericano de IA (ILIA) 2025, CENIA y CEPAL (oct 2025).

Oscar Obando
CHAVES

“Training runs are far easier to conceal than missile silos, their inputs are general-purpose, and the incentive to defect quietly is enormous, because whoever continues while others pause could inherit the lead.”

— Anthropic Institute
"When AI builds itself", 4 de junio de 2026

oscarobando.ai
linkedin.com/in/oscar-obando-chaves-237495133

Junio 2026 · Quito, Ecuador