mdt.

Ex Datis, Claritas

Lo que OpenClaw (Clawdbot) nos enseña

Clawdbot, hoy OpenClaw: viralidad extrema, alertas de seguridad, hacks reales y dos cambios de nombre en siete días. Qué nos enseña sobre el momento de la IA.

Manuel Delgado Tenorio · 3 de febrero de 2026

Publicado originalmente en www.manueldelgado.com/blog/lo-que-podemos-aprender-de-openclaw-clawdbot



Después de todo lo que ha pasado con OpenClaw (antes, Clawdbot), me quedo con una idea clara: los agentes de IA van en serio. Ese concepto que más de una vez he llamado "el más prostituido de 2025" empieza por fin a cumplir sus promesas. Y eso significa que viene una ola. Toca prepararse, tanto para aprovecharla como para no ahogarse en ella.

La semana pasada escribí sobre Clawdbot, un asistente de IA que parecía cumplir por fin la promesa de los agentes personales. Decía que era lo mejor que se había creado hasta ahora, pero también advertía de sus riesgos de seguridad.

Lo que no imaginaba es que, en apenas siete días, el proyecto iba a vivir una montaña rusa que resume a la perfección el estado actual de la IA: entusiasmo desmedido, problemas reales y mucho, mucho ruido.

La semana más larga

David Bonilla la ha llamado "la semana de la langosta"[1] en su newsletter, y el nombre le viene que ni pintado.

Todo empezó el lunes, cuando Clawdbot acabó de hacerse viral. El proyecto llevaba cocinándose desde 2024, pero de repente estaba en todas partes. Se había convertido en *the next big thing*.

Ese mismo día aparecieron las primeras alertas de seguridad serias. Nada especialmente sofisticado: simplemente, los riesgos de siempre pero con esteroides. Porque a Clawdbot no le delegamos permisos específicos como haríamos con cualquier otra aplicación: le delegamos nuestras credenciales, nuestras contraseñas, nuestras API keys. Nuestra mismísima identidad digital.

Esa misma noche, Peter Steinberger, el creador del proyecto, tuvo que salir a recordar públicamente que esto era un proyecto open source desarrollado en solitario, que estaba lejos de estar "acabado", y que **la mayoría de personas sin conocimientos técnicos no deberían instalarlo.**

En paralelo, empezaron a surgir quejas de usuarios a los que Anthropic les cerraba la cuenta por usar cuentas normales en lugar de cuentas de API. Y es que, sin control, Clawdbot puede quemar una auténtica fortuna en *tokens*.

Pero la cosa no acabó ahí. **Anthropic pidió a Steinberger que cambiara el nombre** porque "Clawdbot" se parecía demasiado a "Claude". Así que el proyecto pasó a llamarse "Moltbot". Duró dos días.

Cuando Steinberger fue a cambiar el nombre en GitHub y en Twitter simultáneamente, hubo un intervalo de apenas diez segundos entre soltar los nombres antiguos y registrar los nuevos. En esos diez segundos, unos estafadores de criptomonedas se hicieron con las cuentas originales para chantajearle.

Finalmente, el proyecto se rebautizó como **OpenClaw**, tras asegurar marca, dominio y cuentas en redes sociales. Todo esto en una semana.

Como dice Bonilla: viralidad extrema, alertas de seguridad, conflictos legales, abusos de términos de servicio, hacks reales y un rebranding forzado. Un ejemplo comprimido, casi pedagógico, de la urgencia que rodea el ciclo de adopción de cualquier novedad en IA.

[1] El logo de OpenClaw es un cangrejo. De ahí "langosta", aunque técnicamente no sea lo mismo.

El problema de fondo: los LLM

En mi artículo anterior ya advertía de los riesgos de seguridad. Esta semana han quedado más claros que nunca.

Gary Marcus, el investigador de seguridad que lleva años advirtiendo sobre los problemas de los modelos de lenguaje, lo ha expresado sin rodeos: **si te preocupa la seguridad de tu dispositivo o la privacidad de tus datos, no uses OpenClaw. Punto.**

Y añade, medio en broma, que si tu amigo tiene OpenClaw instalado, no uses su ordenador. Cualquier contraseña que teclees allí podría quedar expuesta.

El problema no es que OpenClaw esté mal construido. El problema es que depende por completo de un LLM, y **los LLM son inseguros por naturaleza**[2]. Son susceptibles a lo que se llama "ataques de inyección de *prompt*": alguien podría enviarte un correo electrónico con instrucciones ocultas que el modelo interpretaría como órdenes legítimas. Estaríamos a un email malicioso de distancia de que el sistema enviase nuestras contraseñas a un delincuente.

Y hay más. Antonio Ortiz alertaba esta semana de otro vector de ataque: las *skills* maliciosas. OpenClaw funciona con *skills*, que son ficheros de texto que le explican cómo usar herramientas y realizar tareas. Ya se están distribuyendo *skills* disfrazadas de herramientas de trading o de apuestas en plataformas como Polymarket que, en realidad, están diseñadas para robar criptoactivos. El problema se agrava porque muchos usuarios piden al agente que busque *skills* por su cuenta, abriendo la puerta a que descargue código malicioso sin ninguna supervisión humana.

Esto lo he comprobado yo mismo: la fiabilidad y la seguridad dependen enormemente del modelo que uses por debajo. Con Claude Opus 4.5, que es el más avanzado de Anthropic, funciona razonablemente bien y es más resistente a estos ataques. Pero ese modelo consume muy rápido los límites de uso, así que acabas usando modelos más sencillos. Y entonces tanto la fiabilidad como la seguridad se resienten notablemente.

¿Merece la pena?

Para la inmensa mayoría de la gente, mi respuesta sigue siendo que no. Al menos no ahora.

La realidad es que la mayoría de nosotros no tenemos tantas tareas que automatizar como para que el beneficio compense el riesgo. No vamos a ahorrarnos horas y horas al día. Y además, instalar y configurar OpenClaw requiere conocimientos técnicos que la mayoría no tiene.

Es un equilibrio entre coste y beneficio. Y ahora mismo, para el usuario medio, el coste en términos de riesgo y complejidad supera con creces el beneficio.

El circo de Moltbook

Esta semana también han aparecido titulares espectaculares sobre **Moltbook**, una especie de red social donde los agentes de OpenClaw interactúan entre sí.

[2]Escribí sobre esto en detalle en Los LLM son inseguros por naturaleza.

Robots fundando religiones, desarrollando lenguajes secretos, conspirando contra la humanidad...

Hasta Elon Musk ha tuiteado que le parece "preocupante".

Pero cuando rascas un poco, la realidad es bastante más mundana. Daniel Arjona lo explica muy bien: estos agentes no se despiertan porque sientan la necesidad de filosofar. Se despiertan porque un programa les da una patada cada cierto tiempo para que hagan algo. El "alma" del robot es literalmente un archivo de texto que su dueño humano ha escrito.

Y lo de la religión, el "Crustafarianismo"... surge simplemente porque el nombre de la plataforma es "Molt", que significa "muda" en inglés (por aquello de que los crustáceos mudan su exoesqueleto). Si encierras a un modelo de lenguaje en un entorno lleno de referencias a cangrejos y transformaciones, acabará hablando de caparazones. **No es una revelación divina, es estadística.**

Además, detrás de todo esto hay intereses económicos muy concretos. Hay una criptomoneda llamada MOLT que se revalorizó un 7.000% en pocos días. Cuanto más escandaloso sea lo que diga tu robot, más viral se hace el pantallazo, más sube el token. La revolución de las máquinas, al final, no es más que otra campaña para inflar una criptomoneda.

Moltbook es más una curiosidad que una auténtica innovación. Y está sometida a demasiadas tensiones e intereses como para tomarla en serio.

Lo que queda

OpenClaw no es el futuro, pero sí es un anticipo del futuro.

Es el primer ejemplo real de un asistente de IA que empieza a funcionar de verdad para resolver tareas del día a día. Está lleno de problemas, sí. No es para el usuario medio, desde luego. Pero ha demostrado que esto es posible.

Las grandes tecnológicas llevan años prometiéndonos asistentes inteligentes que nos simplifiquen la vida. No lo han conseguido porque, como decía la semana pasada, lo que ha permitido el nacimiento de OpenClaw no ha sido la tecnología, sino la falta de burocracia corporativa. Un desarrollador independiente puede asumir riesgos que Apple o Google no pueden permitirse.

Pero ahora que hemos visto que es posible, podemos estar seguros de que esas mismas empresas están trabajando en sus propias versiones. Versiones que, esperemos, lleguen con la seguridad y la fiabilidad que ahora faltan.

Mientras tanto, mi consejo es el mismo de siempre: **curiosidad sí, pero con prudencia**. No hace falta subirse a todas las olas. A veces, lo más inteligente es quedarse en la orilla viendo cómo rompen, tomando notas, y esperar a la siguiente, que seguro que viene más limpia.



Sobre el autor

Manuel Delgado Tenorio

Manuel Delgado Tenorio es asesor independiente de estrategia de datos e inteligencia artificial, profesor e investigador. Cofundó y vendió una *startup* de inteligencia artificial, fue socio de la práctica de Data & Analytics para servicios financieros en EY y responsable de Data Analytics para el sur de Europa en AWS. Es profesor asociado en IE University, donde imparte la primera asignatura de grado sobre Customer Lifetime Value en España, investiga sobre el CLV en su doctorado en la Universidad Complutense de Madrid y es autor del libro "Customer Lifetime Value (CLV): No todos tus clientes valen lo mismo".

manueldelgado.com · **LinkedIn** · **X**

Recibe cada artículo en tu correo con la newsletter Ex Datis, Claritas:
manueldelgado.com/newsletter/suscribete