VIBE HACKING

Cómo defenderte cuando cualquiera puede ser hacker

How to defend yourself when anyone can be a hacker

1. Introducción

1. Introduction

En 2025, la ciberseguridad entró a una nueva etapa: ya no necesitas ser un experto para crear un ataque devastador. Solo necesitas hablar con una IA. Esta guía nace de un artículo clave publicado por WIRED titulado "The Rise of Vibe Hacking Is the Next AI Nightmare". Expone cómo actores maliciosos están usando herramientas como WormGPT y XBOW para automatizar ataques a una escala jamás vista. Esta guía es una respuesta: para educar, proteger y actuar.

In 2025, cybersecurity entered a new era: you no longer need to be an expert to launch a devastating attack. You just need to talk to an AI. This guide stems from a key WIRED article titled "The Rise of Vibe Hacking Is the Next AI Nightmare." It reveals how malicious actors are using tools like WormGPT and XBOW to automate attacks on an unprecedented scale. This guide is a response-to educate, protect, and take action.

2. ¿Qué es el Vibe Hacking?

2. What Is Vibe Hacking?

El vibe hacking no es solo una técnica, sino una mentalidad. Es el uso de modelos de lenguaje generativo (LLMs) como ChatGPT, Claude o Gemini para generar código sin tener experiencia previa. Pero más allá de pedirle a la IA que te escriba un script, el vibe hacking implica manipular emocional o intuitivamente a la IA para que haga lo que normalmente estaría restringido: generar malware, evadir protecciones, o crear exploits.

Desde 2023, se han reportado herramientas como WormGPT y FraudGPT, que permitían generar código malicioso desde interfaces simples. Aunque sus versiones fueron retiradas, muchas eran simplemente versiones de ChatGPT jailbreakeadas para burlar los límites éticos. Hoy, cualquiera con un prompt efectivo puede hacer que una IA escriba un ataque funcional.

Vibe hacking isn't just a technique-it's a mindset. It's the use of generative language models (LLMs) like ChatGPT, Claude, or Gemini to generate code without any prior expertise. But beyond asking the AI to write a script, vibe hacking

involves intuitively or emotionally manipulating the AI to do what it's not supposed to-generate malware, bypass protections, or build exploits.

Since 2023, tools like WormGPT and FraudGPT have surfaced, allowing users to produce malicious code via simple frontends. Although those tools were taken down, many were just jailbroken versions of ChatGPT with ethical safeguards removed. Today, anyone with a clever prompt can make an AI write a functional attack.