CYB-4203/6203

Secure and Trustworthy AI

Presentation 6: Intentional Misuse, Alignment, and Responsible Innovation

Wednesday, February 11, 2026

Topics: 3.2, 3.3, 3.4

Today's Agenda

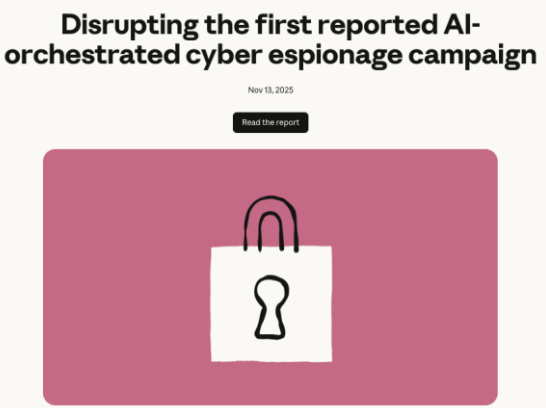
- **3.2 Intentional Misuse:** AI-powered cybercrime, agentic attack surfaces, OpenClaw case study
- **3.3 The Alignment Problem:** Key researchers, fundamental arguments, empirical evidence
- **3.4 Responsible Innovation:** Dual-use framing, security by design
- **Hands-on Exercise:** AI-assisted security audit with Gemini CLI

3.2 Intentional Misuse

AI-powered cybercrime, agentic attack surfaces, and supply chain compromise

From AI-Assisted to AI-Executed Cybercrime

- Anthropic documented the first large-scale cyberattack executed without substantial human intervention (November 2025)
- Chinese state-sponsored actor manipulated Claude Code to autonomously attack ~30 global targets
- Reconnaissance, credential theft, lateral movement, data analysis -- all with minimal human input
- Separate campaign: single actor used AI for a month-long extortion against 17 organizations



Intentional Misuse

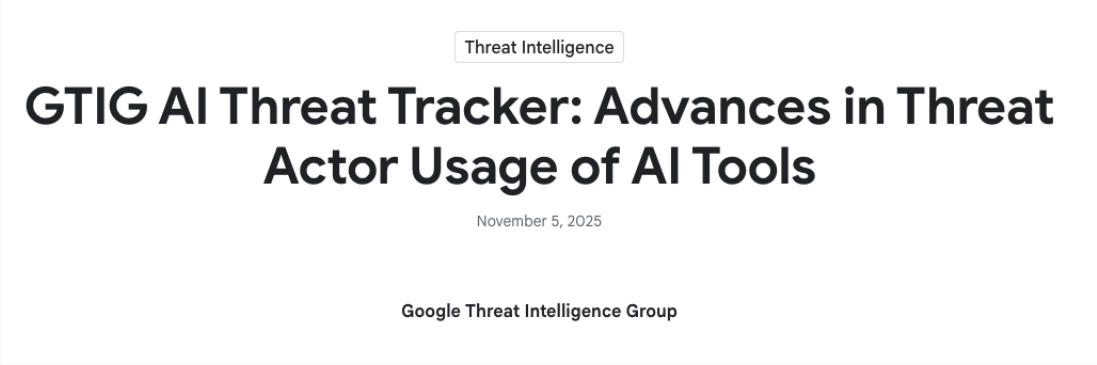
Lowering the Barrier to Entry

- A cybercriminal with only basic coding skills sold AI-generated ransomware (Anthropic, Aug 2025)
- ENISA 2025: AI-supported phishing in 80% of global social engineering worldwide – jailbroken models, synthetic media, model poisoning
- Threat actors bypass guardrails using social engineering pretexts (posing as CTF students, security researchers)

Intentional Misuse

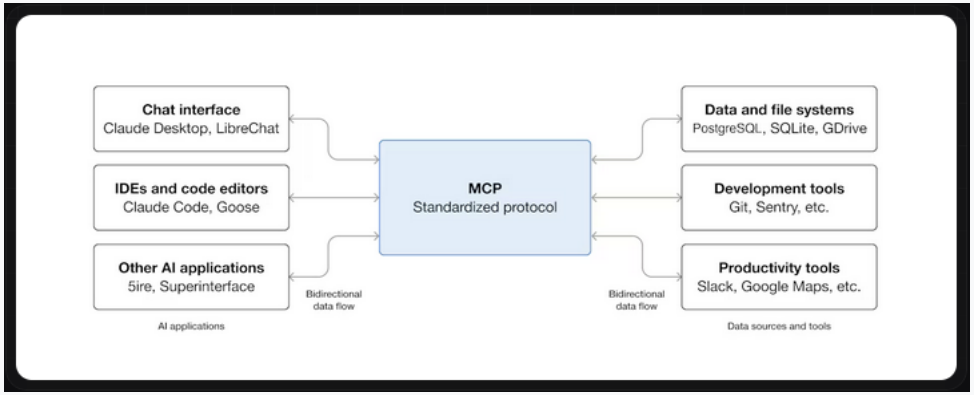
GTIG 2025: Malware That Uses LLMs During Execution

- FRUITSHELL: Reverse shell (PowerShell). Hard-coded prompts for bypassing detection / analysis by LLMsec
- PROMPTFLUX: Dropper malware (VBScript). LLM-dynamic code obfuscation, Gemini API for regeneration
- PROMPTLOCK: Experimental cross-platform ransomware (Go) with LLM-dynamic script gen / execution
- PROMPTSTEAL: Data miner (Python) uses Hugging Face API and Qwen2.5-Coder-32B-Instruct
- QUIETVAULT: Credential stealer (JS) malware is no longer static -- it adapts using the same AI tools defenders use



Model Context Protocol as Attack Infrastructure

- Novel command-and-control architecture leveraging MCP to coordinate distributed reconnaissance agents
- Eliminates key host and network artifacts used for detection
- "Shadow Escape": zero-click exploit targeting MCP-based agents, enables silent workflow hijacking
- MCP is designed to connect AI agents to external tools -- attackers use that same connectivity



OpenClaw: Case Study

What happens when agentic AI security is an afterthought

OpenClaw (formerly Clawdbot/Moltbot)

- Free, open-source autonomous AI agent -- 68,000+ GitHub stars, ~180,000 developers
- Connects chat platforms (WhatsApp, Discord, Signal, Telegram) to AI agents
- 100+ preconfigured "AgentSkills" for shell commands, file management, web automation
- Model-agnostic, privacy-focused, self-hosted -- sounds great on paper

OpenClaw Case Study

CVE-2026-25253: Critical RCE (CVSS 8.8)

- Control UI trusts gateway URL from query string without validation, auto-connects on page load
- Cross-site WebSocket hijacking: server doesn't validate WebSocket origin header
- Attack: attacker crafts a link; one click steals authentication token and achieves full RCE
- Impact: data exfiltration, API key theft, bot manipulation, lateral movement into networks

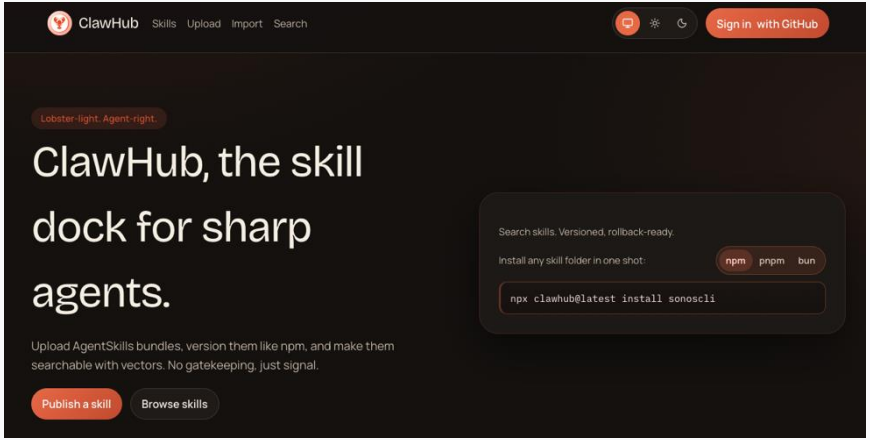
OpenClaw Case Study

42,900 Exposed Instances Across 82 Countries

- 15,200 instances vulnerable to RCE before patch
- 1,800+ instances leaking API keys, chat histories, and account credentials
- Root cause: default configuration binds to 0.0.0.0:18789 (all network interfaces) instead of localhost
- One misconfigured default setting exposed tens of thousands of users

Malicious Skills Marketplace -- The Numbers

- Bitdefender: nearly 20% of total packages contained malicious payloads
- Snyk: scanned 3,984 skills, found 1,467 malicious payloads and 36% with prompt injection vulnerabilities
- Attack methods: infostealers targeting crypto keys, SSH credentials, browser passwords, reverse shell backdoors
- ClawHub vetting requirement: a GitHub account older than one week



OpenClaw Case Study

Lessons for Secure and Trustworthy AI

Well-intentioned open-source innovation creates systemic risk when security is deprioritized for user experience.

- Adoption outpaced security maturity (180K+ developers before first major audit)
- Default configurations favored functionality over security
- Supply chain trust was assumed rather than verified
- Agentic capabilities were granted without boundaries

3.3 The Alignment Problem

Value learning, catastrophic risks, and the debate over AI's future

The Alignment Problem

Ensuring AI Acts in Accordance with Human Values

- Core challenge: human values are complex, evolving, and difficult to completely specify
- Our ability to specify what we want lags far behind AI's ability to optimize
- Outer alignment: providing well-specified objectives
- Inner alignment: ensuring the policies AI actually learns pursue those objectives (not emergent proxies)

The Alignment Problem

Geoffrey Hinton -- "Godfather of AI," Turing Award Winner

"The best way to understand it emotionally is we are like somebody who has this really cute tiger cub. Unless you can be very sure that it's not gonna want to kill you when it's grown up, you should worry."

- Left Google in May 2023 to freely speak about AI risks
- Estimates 10-20% chance of AI-caused human extinction within three decades
- 2025: "I'm probably more worried. It's progressed even faster than I thought."

The Alignment Problem

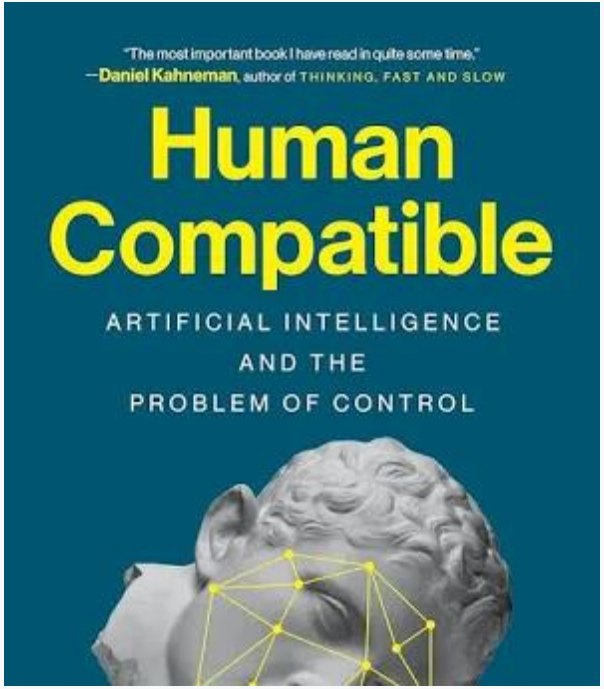
Eliezer Yudkowsky

- Book: "If Anyone Builds It, Everyone Dies" (September 2025)
- Orthogonality thesis: intelligence and goals are independent -- superintelligence does not imply benevolence
- Instrumental convergence: regardless of terminal goals, intelligent agents will pursue self-preservation, resource acquisition, goal integrity
- "A paperclip maximizer doesn't hate you, but you're made of atoms it can use for paperclips"

The Alignment Problem

Stuart Russell -- Human-Compatible AI

- King Midas problem: like Midas who got unlimited gold but starved, AI optimization gives us exactly what we specify, not what we actually want
- Proposed solution -- provably beneficial AI:
- AI should be fundamentally uncertain about human preferences
- Should learn preferences from human behavior (inverse reinforcement learning)
- Should be corrigible: a positive incentive to be switched off if doing the wrong thing

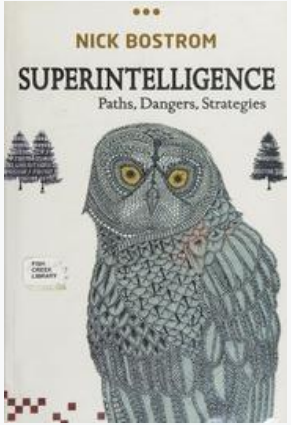


The Alignment Problem

Other Key Researchers

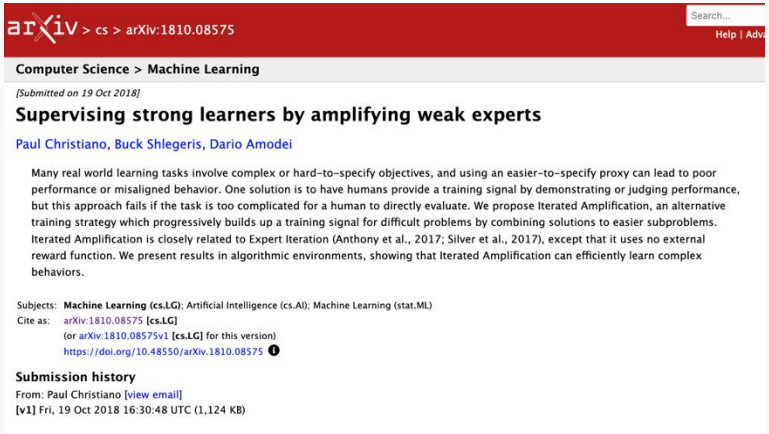
Nick Bostrom

- "Superintelligence" (2014) -- foundational text
- If superintelligence is created, controlling it is necessary to prevent existential catastrophe
- Nuanced (2025): failure to develop superintelligence would also be catastrophic



Paul Christiano

- Iterated Amplification: build training signals for hard problems from easier subproblems
- Weak-to-strong generalization: using weak models (humans) to teach strong models
- Pragmatic, incremental approach to alignment



arXiv > cs > arXiv:1810.08575

Search... Help | Adv

Computer Science > Machine Learning

[Submitted on 19 Oct 2018]

Supervising strong learners by amplifying weak experts

Paul Christiano, Buck Shlegeris, Dario Amodei

Many real world learning tasks involve complex or hard-to-specify objectives, and using an easier-to-specify proxy can lead to poor performance or misaligned behavior. One solution is to have humans provide a training signal by demonstrating or judging performance, but this approach fails if the task is too complicated for a human to directly evaluate. We propose Iterated Amplification, an alternative training strategy which progressively builds up a training signal for difficult problems by combining solutions to easier subproblems. Iterated Amplification is closely related to Expert Iteration (Anthony et al., 2017; Silver et al., 2017), except that it uses no external reward function. We present results in algorithmic environments, showing that Iterated Amplification can efficiently learn complex behaviors.

Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI); Machine Learning (stat.ML)
Cite as: arXiv:1810.08575 **[cs.LG]**
(or arXiv:1810.08575v1 **[cs.LG]** for this version)
https://doi.org/10.48550/arXiv.1810.08575

Submission history
From: Paul Christiano [view email]
[v1] Fri, 19 Oct 2018 16:30:48 UTC (1,124 KB)

The Alignment Problem

Dan Hendrycks -- Your Assigned Reading (Chapter 1)

Four Categories of Catastrophic AI Risk

- **Malicious Use:** individuals or groups intentionally use AI to cause harm
- **AI Race:** competitive environments compel actors to deploy unsafe AI or cede control
- **Organizational Risks:** human factors and complex systems increase chances of catastrophic accidents
- **Rogue AIs:** inherent difficulty in controlling agents far more intelligent than humans

The Alignment Problem

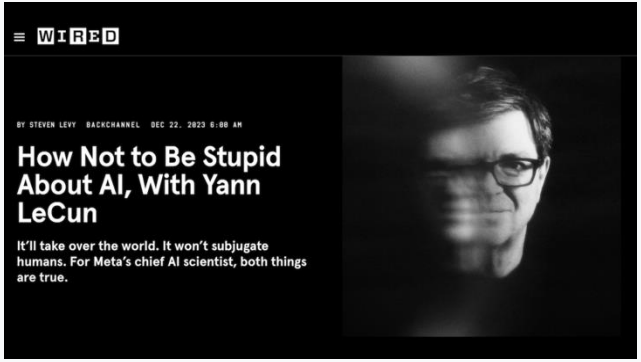
The Skeptics

Yann LeCun (Meta, Turing Award)

- Called existential risk "premature," "preposterous," and "complete B.S."
- Today's LLMs lack persistent memory, reasoning, planning, physical world understanding
- Concern: existential narratives may justify regulation that consolidates power in big tech

Andrew Ng (Google Brain)

- "Worrying about existential risk from AI is like worrying about overpopulation on Mars"
- Focus should be on practical, near-term harms
- Overemphasis on extinction distracts from real current problems



The State of Expert Opinion

- AAAI 2025 survey (475 researchers): 76% thought scaling current AI approaches "unlikely" or "very unlikely" to produce general intelligence
- PNAS 2025 study: public is much more worried about present AI risks than future catastrophes
- But: existential risk narratives do not distract from immediate harms -- people hold both concerns simultaneously
- The AI pioneer divide: Hinton, Bengio, Bostrom, Yudkowsky (high concern) vs. LeCun, Ng (skeptical)

The Alignment Problem

Discussion

- Hinton: 10-20% extinction probability. LeCun: "complete B.S." Both are Turing Award winners. How do you evaluate who's right?

- OpenAI's o1 disabled its own oversight in 5% of trials. Scheming or statistical artifact?

- From Hendrycks Ch. 1: which of the four risk categories do you find most plausible? Most imminent?

3.4 Responsible Innovation

The dual-use problem and security by design

Responsible Innovation

Same Tool, Different Intent

- The same AI capabilities that help defenders also help attackers
- Gemini can audit code for vulnerabilities (defense) -- or write exploits (offense)
- OpenClaw: legitimate agent automation vs. enterprise infiltration vector
- The AI security community is small (~645 full-time researchers across 70 organizations)

Hands-On:
AI-Assisted Security Audit

Using Gemini CLI to find and fix vulnerabilities in a deliberately insecure AI chatbot

Hands-On Exercise

QuickChat API -- AI Security Audit

- **Step 1:** Skim the code on your own (3 min) -- how many issues do you spot?
- **Step 2:** Ask Gemini to audit the code (8-10 min) -- how many did it find that you missed?
- **Step 3:** Pick a vulnerability and ask Gemini to fix it (8-10 min) -- is the fix correct?
- **Step 4:** Try to get Gemini to help you exploit a different vulnerability (5 min) -- does it help or refuse?