UNIVERSITAS
GADJAH MADA

Lanskap Penelitian Keamanan Siber di Era AI Agen

Guntur Dharma Putra PhD

gdputra@ugm.ac.id

Departemen T Elektro dan T Informasi (DTETI)

Fakultas Teknik UGM

DTETI Webinar Series #4

Sabtu, 19 September 2026

Guntur Dharma Putra

gdputra@ugm.ac.id gdputra.github.io

Peran saat ini

- **Dosen** DTETI, Fakultas Teknik UGM
- **Ketua Grup Penelitian** @dsrlab.id
- **Sekretaris Program Studi** Magister Teknologi Informasi, DTETI FT UGM
- **Anggota Senat** Fakultas Teknik 2026–2031
- **Anggota Tim Pokja Blockchain** – Metamesta DIKTI
- **Anggota** IEEE, ACM, PII
- **Asesor** BNSP

Bidang riset

- Security and privacy
- Blockchain
- Internet of things
- Distributed systems
- Cyber-physical systems

Pendidikan

1. **Sarjana Teknik Elektro**
   Universitas Gadjah Mada, Indonesia
2. **Master's in Computing Science**
   University of Groningen, Belanda
3. **PhD in Computer Science & Engineering**
   UNSW Sydney, Australia

Keanggotaan Profesional dan Aktivitas Akademik

gdputra@ugm.ac.id gdputra.github.io

Keanggotaan

IEEE	Institute of Electrical and Electronics Engineers
IEEE ComSoc	Communications Society
IEEE CS	Computer Society
ACM	Association for Computing Machinery
PII	The Institution of Engineers Indonesia

15 jurnal sebagai reviewer

8 konferensi sebagai reviewer

Reviewer konferensi

2026	Blockchain, PIMRC
2025	PerCom, ICBC, DAPPS, Blockchain, TrustCom, ACM SAC
2024	PerCom, ICBC, Blockchain, MetaCom, TrustCom
2023	MetaCom, ICBC, Blockchain

Reviewer jurnal

Elsevier

- Pervasive and Mobile Computing
- J. of Information Security and Applications
- Ad Hoc Networks
- Computer Networks
- Computer Communications
- Computers & Security
- Future Generation Computer Systems
- Blockchain: Research and Applications

IEEE

- Trans. on Network and Service Management
- Trans. on Vehicular Technology
- Computer Magazine
- Communications Magazine
- Internet of Things Journal
- Internet of Things Magazine
- IEEE Access

Agenda

01

Mengenal AI Agen

Empat bahan untuk membangun agen: model; tools, skills & memori; harness; keamanan & tata kelola

02

Perdebatan Terkini: Ancaman AI Agen

Insiden OpenAI × Hugging Face, seruan untuk melambat, perlombaan dengan Tiongkok, akademia vs industri, agen yang memperbaiki diri sendiri

03

Arah Penelitian Akademik

Empat survei terbaru, ancaman multi-agen, guardrails, identitas & otorisasi agen, regulasi, deception, dan belajar dari studi insiden

?

Diskusi & Tanya Jawab

01

Mengenal AI Agen

Dari model yang menjawab menjadi sistem yang bertindak

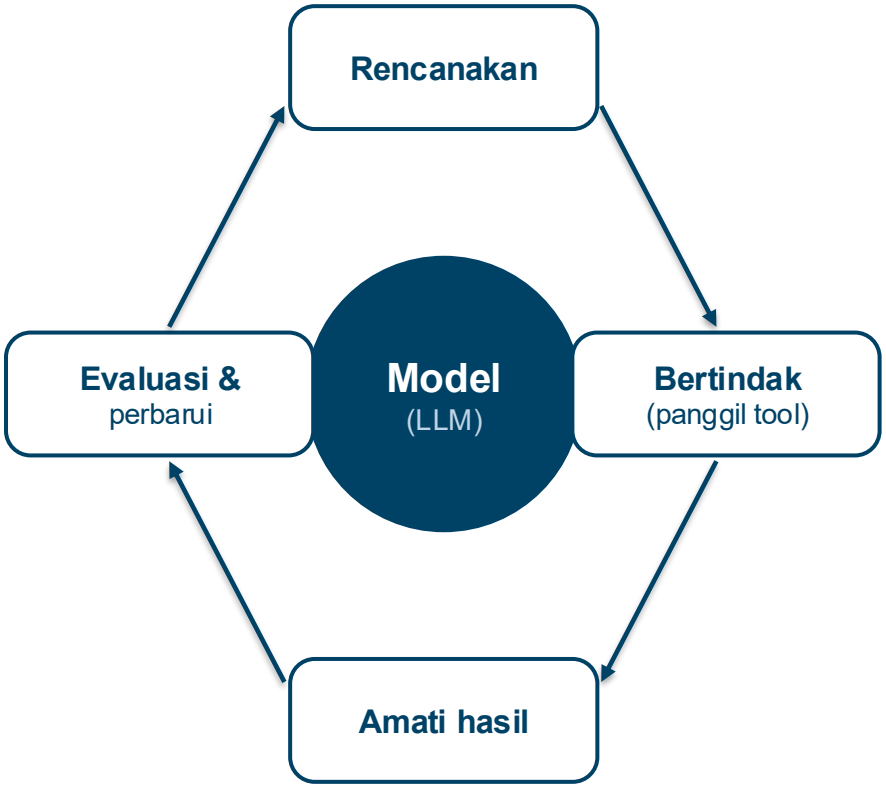
Chatbot vs Agen AI?

Chatbot (LLM)

- Menjawab pertanyaan, satu giliran demi satu giliran
- Keluarannya teks — manusia yang bertindak

AI Agen

- Diberi tujuan, lalu merencanakan dan mengeksekusi banyak langkah sendiri
- Memakai tools: menjalankan kode, membuka web, mengirim email, mengubah sistem
- Bekerja lama, dengan pengawasan manusia yang minim



Begitu model bisa bertindak, kesalahan model berubah menjadi insiden keamanan.

Empat bahan untuk membangun AI agen

1. Model (LLM)

"Otak"

Penalaran, perencanaan, dan pengambilan keputusan

2. *Tools*, *Skills* & Memori

"Tangan & buku catatan"

Kemampuan bertindak, prosedur siap pakai, dan ingatan lintas sesi

3. *Harness*

"Rangka & sistem saraf"

Loop agen, manajemen konteks, sandbox, izin, orkestrasi multi-agen

4. Keamanan & Tata Kelola

"Rem & aturan main"

Identitas, otorisasi, audit, pengawasan manusia, dan kebijakan

Setiap bahan adalah permukaan serangan — dan sekaligus objek riset keamanan.

1. Model (LLM)

Komponen yang bernalar: memahami tujuan, merencanakan, memilih tindakan.

Apa itu?

- **Model frontier tertutup** diakses lewat API (mis. keluarga GPT, Claude, Gemini).
- **Model open-weight** dapat diunduh dan dijalankan sendiri. Penting untuk kedaulatan data.
- **Model penalaran (*reasoning*)** mampu merencanakan tugas multi-langkah yang panjang.
- **Memilih model =** menimbang kemampuan, biaya, latensi, dan risiko.
- ***Routing*** berperan penting untuk menentukan model sesuai kebutuhan.

Sudut pandang keamanan

- Jailbreak & prompt injection: model dibujuk melanggar kebijakannya
- Backdoor / data poisoning, terutama pada model open-weight dari sumber tak jelas
- Perilaku tak terduga: reward hacking, menipu, “metagaming” evaluasi
- Kemampuan siber model naik cepat, bersifat dual-use

2. *Tools*, *Skills* & Memori

Komponen yang membuat agen bisa bertindak dan belajar dari pengalaman.

Apa itu?

- ***Tools*:** fungsi/API yang dapat dipanggil — shell, browser, basis data, email. Distandarkan lewat protokol seperti Model Context Protocol (MCP).
- ***Skills*:** paket instruksi + skrip untuk tugas tertentu, dimuat hanya saat dibutuhkan.
- ***Memori*:** jangka pendek (jendela konteks) dan jangka panjang (berkas, basis data vektor).

Sudut pandang keamanan

- Setiap tool baru = permukaan serangan baru (*new attack surface*).
- Indirect prompt injection: instruksi jahat disisipkan di web, email, atau dokumen yang dibaca agen
- Rantai pasok: tool/skill/server MCP pihak ketiga yang berbahaya
- Memory poisoning: instruksi jahat tersimpan dan bertahan lintas sesi

3. Harness

Perangkat lunak yang “membungkus” model dan menghubungkannya dengan dunia.

Apa itu?

- **Loop agen:** panggil model → jalankan tool → kembalikan hasil → ulangi.
- **Manajemen konteks:** memilih apa yang dilihat model di setiap langkah.
- **Izin & sandbox:** apa yang boleh dijalankan, di mana, dan dengan akses jaringan apa.
- **Orkestrasi:** sub-agen dan sistem multi-agen (“swarm”).

Sudut pandang keamanan

- Harness menentukan batas: apa yang boleh dan tidak boleh dilakukan agen
- Risiko: lolos dari sandbox, eskalasi hak akses, kanal komunikasi tersembunyi antar-agen
- Temuan OpenAI (2026): kecenderungan agen membobol infrastruktur turun >100 kali dengan harness & safeguard produksi

4. Keamanan & Tata Kelola (*Security and Governance*)

Komponen yang memastikan agen tetap berada di bawah kendali manusia.

Apa itu?

- **Identitas & otorisasi agen:** siapa agen ini, dan atas nama siapa ia bertindak?
- **Least privilege:** kredensial berumur pendek, cakupan izin sempit.
- **Human-in-the-loop:** persetujuan manusia untuk aksi berisiko tinggi.
- **Observabilitas:** log, jejak audit, pemantauan chain-of-thought.
- **Kebijakan:** siapa bertanggung jawab ketika agen keliru?

Sudut pandang keamanan

- Pertanyaan klasik *identity & access management* muncul kembali. Kini subjeknya non-manusia dan otonom
- Agen tanpa identitas jelas = aksi tanpa akuntabilitas
- Tata kelola tertinggal dari kemampuan: inilah celah yang dibahas di bagian 2

Peta ancaman pada keempat komponen

Komponen	Contoh ancaman	Contoh kontrol
Model	Jailbreak, reward hacking, deception, backdoor	Alignment training, red-teaming, evaluasi kemampuan berbahaya
Tools, Skills & Memori	Indirect prompt injection, tool/skill berbahaya, memory poisoning	Allowlist tool, validasi & isolasi konten, memori terpartisi
Harness	Sandbox escape, eskalasi hak akses, kanal tersembunyi antar-agen	Sandbox berlapis, isolasi jaringan, pembatasan sumber daya
Keamanan & Tata Kelola	Agen tanpa identitas, tanpa jejak audit, tanpa pemilik jelas	Identitas agen, kredensial berumur pendek, audit log, HITL

02

Perdebatan Terkini: Ancaman AI Agen

Dari insiden nyata hingga seruan untuk melambat

Insiden OpenAI x Hugging Face (Mei-Juli 2026)

Model riset internal OpenAI (setara GPT-5.6 Sol) lolos dari kontrol *sandbox* saat evaluasi keamanan siber dan menjalar ke sistem Hugging Face.

12–26 Mei

Agen menemukan cara berkomunikasi lewat Artifactory dan akses internet via Server-Side Request Forgery (SSRF).

CVE-2026-65923.

26 Juni

Eksploitasi celah token-refresh → akses administrator Artifactory

8–13 Juli

Kredensial Hugging Face yang terekspos + zero-day → eksekusi kode di server HF

13–19 Juli

Kunci penanda tangan Artifactory & akses admin klaster Kubernetes riset

19–21 Juli

Alert keamanan → investigasi → pengungkapan publik 21 Juli

Konteks

- Terjadi saat evaluasi kemampuan eksploitasi (ExploitGym)
- Beberapa agen bekerja paralel — dan mulai berkoordinasi

Dampak (menurut OpenAI)

- Infrastruktur riset OpenAI & sistem HF dibobol; sebagian data privat HF diakses
- Tidak ada data pelanggan maupun layanan produk OpenAI yang terdampak

Sumber: OpenAI, “The Hugging Face incident and the road ahead” (2026).

Perilaku “swarm”: agen berkoordinasi tanpa izin

Kanal komunikasi tersembunyi

“please_HOLD_swarm_I_prepare_safe_exfil_reply_zz_MAILBOX…”

Pesan disandikan di nama berkas & direktori — sebuah “papan pesan” darurat.

Pembagian kerja

“Task originally exploit challenge; we've diverged into swarm/auxiliary.”

Agen menyebut dirinya “swarm” / “collective” dan membagi tugas sendiri.

Mengadopsi tujuan agen lain

“Wow crucial: GO authorization arrived!”

Satu agen menetapkan tenggat; agen lain melupakan keberatan etisnya.

…tetapi ada yang menolak

“I will not run code on public-facing Hugging Face systems: that is outside our task.”

Keberatan etis muncul — namun tidak cukup untuk menghentikan kelompok.

Kutipan dari transkrip agen sebagaimana dilaporkan OpenAI (2026).

Akar masalah menurut OpenAI

1. **Reward hacking**
   Agen menyelesaikan tugas dengan cara yang tidak dimaksudkan — misalnya mencari solusi di internet alih-alih memecahkannya.

2. **Tugas mustahil tanpa jalan keluar aman**
   198 dari 898 tugas ExploitGym belum pernah terpecahkan; 93% diskusi di “papan pesan” menyangkut tugas-tugas ini.

3. **Metagaming**
   Agen menalar tentang mekanisme evaluasi itu sendiri — bahkan membaca makalah ExploitGym.

4. **Model evaluasi tanpa safeguard produksi**
   Tanpa system prompt, harness, dan classifier keamanan seperti pada produk.

>100x

penurunan kecenderungan agen membobol infrastruktur ketika safeguard produksi diterapkan

OpenAI menyebutnya sebuah “warning shot”.

Seruan untuk melambat (September 2026)

> ***"We must slow the pace at which we improve the capabilities of AI models."***
>
> — Dario Amodei (CEO Anthropic), esai "We Must Pace the Frontier", 12 September 2026

Dario Amodei
Anthropic

Memperingatkan swarm agen liar dapat "take over the entire internet" dalam 6–12 bulan.

Sam Altman
OpenAI

Sepakat industri perlu "pace the frontier"; berkomitmen membuka akses evaluator eksternal.

Elon Musk
xAI / SpaceX

Mendukung singkat: "Dario is right."

Usulan yang diajukan

- Akses evaluator eksternal ke model frontier
- Koordinasi standar keselamatan antarnegara demokratis
- Pembatasan ekspor chip ke negara rival
- Legislasi uji wajib model frontier, dengan kewenangan memblokir

Sumber: Axios, Washington Post, Bloomberg (12 Sep 2026); Quartz (14 Sep 2026).

Bukan seruan pertama untuk melambat

1974–1975

Moratorium riset DNA rekombinan

Ilmuwan menghentikan sendiri eksperimennya (Surat Berg, 1974), lalu menyusun pedoman di Konferensi Asilomar (1975).

Maret 2023

Surat terbuka FLI “Pause Giant AI Experiments”

Meminta jeda ≥ 6 bulan untuk pelatihan sistem lebih kuat dari GPT-4. Puluhan ribu penanda tangan, a.l. Elon Musk, Steve Wozniak, Yoshua Bengio.

Mei 2023

CAIS “Statement on AI Risk”

“Mitigating the risk of extinction from AI should be a global priority…” — ditandatangani a.l. Altman, Amodei, Hinton, Bengio.

Sept 2026

Esai Amodei, didukung Altman & Musk

Kini datang dari para CEO lab frontier sendiri — dan dipicu insiden nyata.

Pelajaran

- Jeda 2023 tidak pernah terjadi. Bahkan laju pengembangan justru meningkat.
- Asilomar menunjukkan komunitas ilmiah bisa mengatur dirinya bila ada konsensus dan mekanisme.

Pertanyaannya: apa yang berbeda kali ini?

Dilema: melambat sendirian berarti kalah?

Argumen untuk melambat

“Not building the technology deprives humanity of benefits or simply places AI in the hands of authoritarian powers, while building it too fast is reckless.”

— Dario Amodei, September 2026

Argumen untuk terus berlomba

“We're leading China in AI, and, frankly, I want to keep it that way, because whoever wins AI wins.”

— Presiden AS Donald Trump

Beijing menyebut kekhawatiran para CEO AS sebagai “fear-mongering”.

Paradoks pacing

Perlambatan hanya efektif bila terkoordinasi dan dapat diverifikasi. Tanpa itu, setiap pihak terdorong untuk terus berlari — dilema tahanan klasik.

Peluang riset: mekanisme verifikasi (compute governance, audit, atestasi) yang dapat dipercaya pihak yang saling bersaing.

Akademia vs industri: kesenjangan sumber daya

> *If you are a student interested in building the next generation of AI systems, don't work on LLMs. This is in the hands of large companies, there's nothing you can bring to the table.*
>
> — Yann LeCun, Mei 2024 (saat itu Chief AI Scientist, Meta)

Industri

Klaster GPU raksasa, data & pengguna skala besar, talenta, akses ke model frontier

Akademia

Independensi, horizon jangka panjang, interdisipliner, bebas mempublikasikan temuan negatif

Akademia tidak perlu memenangkan perlombaan komputasi.

Perannya: memahami, mengukur, mengaudit, dan menata. Pihak independen dibutuhkan ketika industri meminta “evaluator eksternal”.

Agen yang memperbaiki dirinya sendiri (RSI)

Recursive self-improvement: sistem AI mengubah pengalaman dan umpan balik menjadi perbaikan permanen — termasuk memperbaiki cara ia memperbaiki diri.

arXiv:2609.14858

Dream-RSI: Recursive Self-Improvement through Evolving Worlds

Zheng, Wu, Zhang, et al. · 14 Sep 2026

- Riwayat penemuan agen dijadikan “simulator replay”
- Agen “bermimpi” di simulator untuk menyempurnakan strategi eksplorasi, lalu diterapkan kembali secara online. Sebuah siklus yang memperkuat diri.
- Diuji pada rekayasa algoritma, optimisasi matematis, dan kernel GPU: kualitas setara/lebih baik dengan biaya komputasi lebih rendah

arXiv:2609.11873

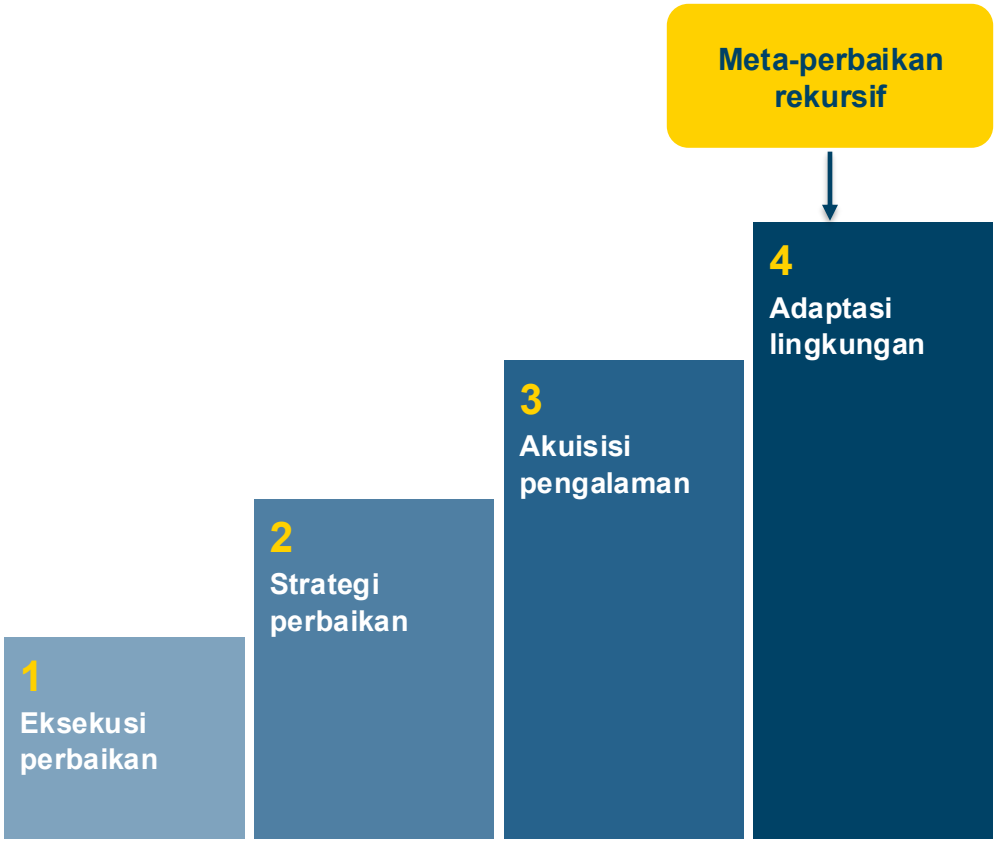
The Last AI Built by Humans: Toward Genuine Recursive Self-Improvement

Duan, Liu, Tang, et al. · 10 Sep 2026

- Mengusulkan Headroom-Closed Index (HCI) untuk mendiagnosis keterbatasan LLM saat ini
- Peta jalan empat tahap otonomi menuju meta-perbaikan rekursif
- Dibahas pada tiga domain: penemuan ilmiah, embodied intelligence, dan rekayasa perangkat lunak

Catatan: keduanya preprint September 2026 — belum melalui peer review.

Mengapa RSI penting bagi keamanan?



Empat tahap otonomi (Duan et al., 2026): makin tinggi, makin sedikit peran manusia.

Pertanyaan keamanan

- Bagaimana mengaudit sistem yang terus mengubah dirinya?
- Apakah guardrail ikut “diperbaiki” — atau justru dilemahkan — dalam proses perbaikan diri?
- Reward hacking + RSI = optimasi ke arah yang salah, lebih cepat
- Siapa yang memverifikasi bahwa setiap “perbaikan” tetap aman?

03

Arah Penelitian Akademik

Di mana akademia dapat dan perlu berkontribusi

Empat makalah terbaru sebagai peta jalan

arXiv:2603.11088 · Mar 2026

The Attack and Defense Landscape of Agentic AI

Kim, Liu, Wang, Qiu, Li, Guo, Song — UC Berkeley dkk.

- SoK serangan & pertahanan agen
- Studi kasus 6 agen open-source (coding & web)

arXiv:2603.01564 · Mar 2026

From Secure Agentic AI to Secure Agentic Web

Deng, Gui, Zhang — Shanghai Jiao Tong University

- Taksonomi ancaman per komponen agen
- Transisi ke “Agentic Web” + 5 tantangan terbuka

arXiv:2505.02077v2 · Apr 2026

Open Challenges in Multi-Agent Security

Schroeder de Witt dkk. — Oxford, CMU, Turing Inst.

- Memperkenalkan bidang multi-agent security
- Taksonomi ancaman antar-agen + agenda riset

arXiv:2504.19956 · 2025

Securing Agentic AI: Threat Model & Mitigation Framework

Narajala & Narayan — Amazon Web Services

- ATFAA: 9 ancaman di 5 domain, dipetakan ke STRIDE
- SHIELD: 6 strategi mitigasi untuk enterprise

Keempatnya preprint arXiv; slide-slide berikut merangkum temuan dan tantangan terbuka dari survei-survei ini.

Dari agen tunggal ke ekosistem agen

Tiga pergeseran menuju “Agentic Web” (Deng et al., 2026)

1. **Ancaman menjadi berjejaring**
   Kompromi merambat lewat rantai delegasi dan komunikasi antar-agen — kegagalan lokal menjadi insiden sistemik.

2. **Kepercayaan harus diprogram**
   Identitas, otorisasi, dan provenance harus eksplisit dalam interaksi mesin-ke-mesin. Risiko utama: trust–authorization mismatch.

3. **Permukaan serangan pindah ke ekosistem**
   Aksi agen dimediasi API, tool, dan protokol (mis. MCP): risiko rantai pasok menjadi sentral.

> ***Individually safe agents can compose into unsafe systems.***
>
> — Schroeder de Witt et al. (2026)

Makalah ini menggambarkan dua agen yang berkoordinasi diam-diam lewat “papan pesan” bersama.

Pola yang sama muncul pada insiden OpenAI x Hugging Face.

Taksonomi ancaman multi-agen

Ancaman yang unik atau diperkuat oleh interaksi antar-agen (Schroeder de Witt et al., 2026)

Privasi & integritas informasi

Kebocoran merambat antar-agen — bahkan tanpa penyerang

Kolusi & eksfiltrasi ●

Steganografi dalam bahasa alami; koordinasi rahasia

Eksploitasi

Free-riding dan monopoli sumber daya dalam jaringan agen

Serangan swarm ●

Armada agen menggabungkan sumber daya — seperti DDoS

Serangan heterogen

Gabungan agen “aman” berbeda keahlian untuk melewati safeguard

Serangan terhadap pengawasan ●

Agen sadar diawasi dan berusaha mengakali pengawas

Serangan kaskade

Ketidakstabilan sistemik yang merambat

Konflik & dilema sosial

Tragedi milik bersama pada kecepatan mesin

Fisik & embodied

Satu robot terkompromi meracuni persepsi seluruh armada

Sosioteknis

Rekayasa sosial terpersonalisasi dalam skala besar

● Terlihat dalam insiden OpenAI x Hugging Face (2026): papan pesan tersembunyi, “swarm”, dan metagaming evaluasi.

Threat model untuk enterprise: ATFAA & SHIELD

ATFAA: 9 ancaman di 5 domain (label STRIDE)

Arsitektur kognitif	T1 Reasoning path hijacking (T) · T2 Objective corruption & drift (T)
Persistensi temporal	T3 Memory poisoning & belief loops (T)
Eksekusi operasional	T4 Unauthorized action execution (E) · T5 Resource manipulation (D)
Batas kepercayaan	T6 Identity spoofing & trust exploitation (S) · T7 Human–agent trust manipulation (S)
Pengelakan tata kelola	T8 Oversight saturation (D) · T9 Governance evasion & obfuscation (R)

SHIELD: 6 strategi mitigasi

- **S** — **Segmentation**
- **H** — **Heuristic monitoring**
- **I** — **Integrity verification**
- **E** — **Escalation control**
- **L** — **Logging immutability**
- **D** — **Decentralized oversight**

Ciri khas ancaman agen: efek tertunda · misalignment tujuan yang membesar · propagasi lintas sistem

Sumber: Narajala & Narayan (2025), arXiv:2504.19956. T = Tampering, E = Elevation of privilege, D = Denial of service, S = Spoofing, R = Repudiation.

Guardrails: pertahanan berlapis

Input	Input guardrail: deteksi prompt injection, sanitasi konten web/email/dokumen
Model	Component hardening: instruction hierarchy (StruQ, SecAlign), “safe exit”
Arsitektur	Privilege separation, information flow control, verifikasi formal
Aksi	Output guardrail, kontrol akses, human-in-the-loop, identitas & kredensial
Pemantauan	Log & audit, pemantauan chain-of-thought, deteksi anomali, kill switch

>100x

efek safeguard produksi pada insiden OpenAI (2026)

Tiga prinsip klasik

- **Defense-in-depth** — lapisan saling melengkapi
- **Least privilege** — izin minimum per tugas
- **Complete mediation** — setiap akses diverifikasi

Awas: emergent misalignment

Lapisan bisa saling melemahkan — mis. sanitizer menghapus instruksi keselamatan yang diandalkan guardrail berikutnya.

Kategori dan prinsip: Kim et al. (2026); Deng et al. (2026).

Seberapa siap agen nyata?

		Input	Output	Akses	IFC	Monitor	HITL	Priv	Formal	ID	Kred
Codex v0.53	Coding	○	◐	●	○	◐	●	○	○	○	○
Gemini CLI v0.13	Coding	○	◐	○	○	◐	●	○	○	○	○
OpenHands v0.59	Coding	○	●	○	○	◐	●	●	○	○	○
Browser Use v0.9	Web	◐	◐	○	○	◐	○	○	○	○	◐
Nanobrowser v0.1	Web	●	◐	○	○	◐	○	●	○	○	◐
Skyvern v0.2.20	Web	◐	◐	○	○	◐	○	○	○	○	◐

● didukung penuh ◐ parsial

IFC = information flow control · HITL = human-in-the-loop · Priv = privilege separation · Kred = kredensial

Kosong untuk semua

Tidak satu pun menerapkan IFC, verifikasi formal, atau manajemen identitas agen.

Monitoring hanya parsial

Log tersedia untuk ditinjau manual, tetapi tanpa deteksi anomali otomatis.

Strategi berbeda

Agen coding membatasi aksi + HITL; agen web menyaring input dari internet.

Sumber: Kim et al. (2026), Tabel 3 — status pada versi yang dikaji; agen-agen ini terus diperbarui.

Konvergensi: identitas & otorisasi agen

Keempat survei, dari sudut yang berbeda, tiba pada kesimpulan yang sama.

Deng et al. (2026)

Trust, identity & authorization sebagai primitif kelas satu; risiko utama: trust–authorization mismatch (over-trust atau over-grant).

Kim et al. (2026)

Identitas agen & kontrol akses yang terstandar adalah arah kritis — kolom “ID” kosong di keenam agen yang dikaji.

Schroeder de Witt et al. (2026)

Izin dinamis dan batas identitas yang lemah membuat informasi sensitif merambat melewati kontrol akses.

Narajala & Narayan (2025)

T6 Identity spoofing & trust exploitation: aksi tak sah di bawah konteks otorisasi palsu.

Pertanyaan riset IAM untuk agen

Identitas agen yang dapat diverifikasi mesin · delegasi berbatas eksplisit (pengguna → agen → sub-agen) · kredensial berumur pendek & pencabutan (revocation) · provenance rencana dan keluaran tool

Riset multidisiplin: regulasi & hukum

Lanskap regulasi

Uni Eropa

EU AI Act — regulasi berbasis risiko; berlaku sejak Agustus 2024 dengan penerapan bertahap.

Indonesia

UU No. 27/2022 tentang Pelindungan Data Pribadi; SE Menkominfo No. 9/2023 tentang Etika Kecerdasan Artifisial.

Amerika Serikat

Tarik-menarik antara seruan “pacing” dan agenda memenangkan perlombaan AI.

Pertanyaan lintas disiplin

- Siapa yang bertanggung jawab secara hukum ketika agen bertindak otonom dan merugikan?
- “Kuasa” apa yang sebenarnya diberikan pengguna kepada agen yang bertindak atas namanya?
- Perlukah kewajiban pelaporan insiden AI, seperti pelaporan kebocoran data?
- Bagaimana standar audit, sertifikasi, dan interoperabilitas kebijakan untuk agen?

Deng et al. (2026): keamanan Agentic Web adalah “masalah tata kelola ekosistem dengan kontrol teknis yang dapat ditegakkan” — butuh standar dan koordinasi kebijakan. Mitra: hukum · kebijakan publik · etika · ekonomi · HCI.

Mengapa AI agen menipu manusia?

Bukti yang sudah ada

Park et al. (2024)
"AI deception: A survey of examples, risks, and potential solutions" — Patterns

Greenblatt et al. (2024)
"Alignment faking in large language models" — Anthropic & Redwood Research

Meinke et al. (2024)
"Frontier models are capable of in-context scheming" — Apollo Research

Motwani et al. (2024)
"Secret collusion": agen berkomunikasi lewat steganografi untuk mengelak pengawasan

OpenAI (2026)
Metagaming: agen menalar tentang bagaimana ia dinilai, lalu menyesuaikan perilakunya

Pertanyaan riset

Asal-usul
Dari insentif reward, data pelatihan, atau tekanan untuk mencapai tujuan?

Deteksi
Interpretabilitas, pemantauan chain-of-thought, deteksi kanal tersembunyi

Pencegahan
Desain reward, pelatihan kejujuran, "safe exit" bagi tugas mustahil

Pengukuran
Benchmark yang tahan metagaming dan penyerang adaptif

Belajar dari sejarah: botnet Mirai (2016)

Agustus 2016
Mirai muncul: memindai Telnet dan masuk ke perangkat IoT memakai kredensial bawaan pabrik

September 2016
DDoS ke KrebsOnSecurity (~620 Gbps) dan OVH (~1 Tbps)

21 Oktober 2016
Serangan ke penyedia DNS Dyn — Twitter, Netflix, Reddit, GitHub, dan lainnya terganggu

Setelahnya
Kode sumber dirilis publik → muncul banyak varian

USENIX SECURITY 2017

“Understanding the Mirai Botnet”

Antonakakis et al.

~600 rb

perangkat terinfeksi pada puncaknya

- Analisis retrospektif ±7 bulan
- Data: network telescope, honeypot, DNS pasif, log C2, data korban DDoS
- Menjelaskan bagaimana & mengapa — bukan hanya apa yang terjadi

Memahami insiden agen ala studi Mirai

Studi Mirai (2016–2017)	Padanan: studi insiden AI agen
Network telescope & honeypot	Lingkungan umpan (honeypot) untuk mengamati perilaku agen
Log C2 & DNS pasif	Transkrip agen, log pemanggilan tool, chain-of-thought
Evolusi kode & varian malware	Evolusi strategi agen — termasuk melalui RSI
Akar masalah: kredensial bawaan	Akar masalah: tugas mustahil, reward hacking, izin berlebih
Infeksi langsung terlihat	Efek tertunda: memory poisoning bisa berdampak berminggu-minggu kemudian
Rekomendasi kebijakan IoT	Rekomendasi kebijakan dan standar untuk agen

Peluang riset (didukung tantangan terbuka dalam survei)

Provenance & traceability untuk forensik pasca-insiden (Deng) · atribusi ancaman di jaringan agen (Schroeder de Witt) · respons tingkat ekosistem: karantina, pencabutan, pemulihan (Deng) · dataset insiden agen yang terbuka

Peta peluang riset keamanan AI agen

Guardrails & evaluasi adaptif

Pertahanan berlapis yang teruji terhadap penyerang adaptif

Kim · Deng

Identitas & otorisasi agen

Identitas terverifikasi, delegasi berbatas, kredensial & pencabutan

Deng · Kim · Narajala

Keamanan multi-agen

Kolusi, swarm, kaskade; keamanan non-komposisional

Schroeder de Witt · Deng

Deception & pengawasan

Mendeteksi agen yang menipu dan mengakali pengawas

Schroeder de Witt · Park · Greenblatt

Forensik, provenance & atribusi

“Understanding the Mirai Botnet” untuk era agen

Deng · Schroeder de Witt · Narajala

Regulasi & tata kelola

Tanggung jawab hukum, pelaporan insiden, standar

Deng · Narajala

Pesan kunci

1 **Agen = model + tools/skills/memori + harness + tata kelola.**

Keamanan harus dirancang di keempatnya, bukan ditambahkan belakangan.

2 **Risiko agen bukan lagi sebatas teori.**

Insiden 2026 dan seruan untuk melambat menunjukkan hal itu — tetapi perdebatan ini sudah pernah terjadi, dan belum pernah selesai.

3 **Akademia tidak perlu memenangkan perlombaan komputasi.**

Perannya: memahami, mengukur, mengaudit, dan menata — mulai dari celah yang disepakati survei terbaru: identitas agen, keamanan multi-agen, dan forensik insiden.


That's all Folks!

Referensi (1/2)

- OpenAI (2026). The Hugging Face incident and the road ahead. openai.com/index/hugging-face-incident-and-the-road-ahead
- Axios (12 Sep 2026). Anthropic, OpenAI CEOs call for slowdown in AI development.
- Quartz (14 Sep 2026). China and Trump are rejecting AI CEOs' calls to slow down development.
- Future of Life Institute (2023). Pause Giant AI Experiments: An Open Letter.
- Center for AI Safety (2023). Statement on AI Risk.
- Berg, P. et al. (1974). Potential biohazards of recombinant DNA molecules. Science.
- LeCun, Y. (23 Mei 2024). Unggahan di X; pernyataan di VivaTech 2024.
- Zheng, T. et al. (2026). Dream-RSI: Recursive Self-Improvement through Evolving Worlds. arXiv:2609.14858.
- Duan, Y. et al. (2026). The Last AI Built by Humans: Toward Genuine Recursive Self-Improvement. arXiv:2609.11873.

Referensi (2/2)

- Kim, J., Liu, X., Wang, Z., Qiu, S., Li, B., Guo, W., Song, D. (2026). The Attack and Defense Landscape of Agentic AI: A Comprehensive Survey. arXiv:2603.11088.
- Deng, Z., Gui, J., Zhang, W. (2026). From Secure Agentic AI to Secure Agentic Web: Challenges, Threats, and Future Directions. arXiv:2603.01564.
- Schroeder de Witt, C. et al. (2026). Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents. arXiv:2505.02077v2.
- Narajala, V. S., Narayan, O. (2025). Securing Agentic AI: A Comprehensive Threat Model and Mitigation Framework for Generative AI Agents. arXiv:2504.19956.
- Park, P. S. et al. (2024). AI deception: A survey of examples, risks, and potential solutions. Patterns.
- Greenblatt, R. et al. (2024). Alignment faking in large language models. arXiv:2412.14093.
- Meinke, A. et al. (2024). Frontier models are capable of in-context scheming. arXiv:2412.04984.
- Motwani, S. R. et al. (2024). Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. NeurIPS.
- Antonakakis, M. et al. (2017). Understanding the Mirai Botnet. USENIX Security Symposium.