Repeater : Publikasi Teknik Informatika dan Jaringan
Volume. 4 Nomor. 3 Juli 2026
e-ISSN : 3046-7276, p-ISSN : 3046-7284, Hal 33-44
DOI: https://doi.org/10.62951/repeater.v4i3.929
Tersedia: https://journal.aptii.or.id/index.php/Repeater

Analisis Etika Profesional terhadap Risiko Keamanan pada Sistem OpenClaw

Muhammad Faiqul Umam Dzunnuroeni[1*], Muhammad Zaky Ramadhan[2], Evy Nurmiati[3]
[1-3]Sistem Informasi, Universitas Islam Negeri Syarif Hidayatullah, Indonesia
[*]*Penulis Korespondensi: muhammad.faiqul24@mhs.uinjkt.ac.id*

Abstract. *This study aims to evaluate the implementation of the built-in security model in a local installation of the OpenClaw autonomous agent and to analyze its implications for information technology (IT) professional ethics. The study employed a qualitative method using a single-case study approach. Primary data were collected through direct simulation experiments comparing secure and insecure configurations within an isolated environment. The results indicate that OpenClaw incorporates native security controls based on a single trust boundary and provides adequate security auditing capabilities. The system also includes data protection mechanisms that minimize the risk of misuse when operated according to the recommended configuration. However, configuring the network to be publicly accessible (0.0.0.0) without authentication violates the principle of least privilege and significantly increases the risk of unauthorized access. These findings demonstrate that vulnerabilities in self-hosted autonomous agents are not solely attributable to technical weaknesses in the system but also reflect failures in adhering to IT professional ethics. Therefore, responsibility is clearly shared between developers and IT practitioners. Developers are responsible for providing adequate security mechanisms, while IT professionals have an ethical obligation to maintain system security, protect user privacy, and ensure that security configurations are implemented in accordance with established best practices.*

Keywords: *Autonomous Agents; Cybersecurity; Least Privilege; OpenClaw; Professional Ethics*.

Abstrak. Penelitian ini bertujuan mengevaluasi penerapan model keamanan bawaan pada instalasi lokal agen otonom OpenClaw serta menganalisis dampaknya terhadap etika profesional teknologi informasi (TI). Penelitian menggunakan metode kualitatif dengan pendekatan studi kasus tunggal. Data primer diperoleh melalui eksperimen simulasi langsung yang membandingkan konfigurasi aman dan konfigurasi berisiko dalam lingkungan yang terisolasi. Hasil penelitian menunjukkan bahwa OpenClaw telah memiliki kontrol keamanan asali yang dibangun berdasarkan batas kepercayaan tunggal (single trust boundary) serta dilengkapi dengan fitur audit keamanan yang memadai. Sistem juga menyediakan mekanisme perlindungan terhadap data sehingga mampu meminimalkan potensi penyalahgunaan apabila dioperasikan sesuai konfigurasi yang direkomendasikan. Namun, tindakan praktisi TI yang melonggarkan konfigurasi jaringan menjadi terbuka (0.0.0.0) tanpa menerapkan autentikasi merupakan pelanggaran terhadap prinsip hak istimewa terkecil (*least privilege*) dan meningkatkan risiko akses tidak sah. Temuan ini menunjukkan bahwa kerentanan pada agen otonom self-hosted tidak semata-mata disebabkan oleh kelemahan teknis sistem, melainkan juga mencerminkan kelalaian dalam penerapan etika profesional TI. Oleh karena itu, terdapat pembagian tanggung jawab yang jelas antara pengembang dan praktisi TI. Pengembang bertanggung jawab menyediakan mekanisme perlindungan yang memadai, sedangkan profesional TI memiliki tanggung jawab etis untuk mempertahankan keamanan sistem, melindungi privasi pengguna, serta memastikan konfigurasi keamanan diterapkan sesuai praktik terbaik.

Kata kunci: Agen Otonom; Etika Profesional; Hak Istimewa Terkecil; Keamanan Siber; OpenClaw.

1. LATAR BELAKANG

Perkembangan pesat kecerdasan buatan (AI) telah mendorong transisi paradigma dari model bahasa besar (LLM) yang murni bersifat percakapan menjadi agen otonom (*autonomous agents*) yang berorientasi pada eksekusi tugas (Wang et al., 2026). Berbeda dengan asisten virtual konvensional, agen AI modern dibekali dengan kemampuan untuk mengeksekusi fungsi sistem operasi, memanipulasi file lokal, dan berinteraksi langsung dengan antarmuka eksternal.

Naskah Masuk: 15 April Mei 2026; Revisi: 17 Mei 2026; Diterima: 1 Juni 2026; Tersedia: 08 Juli 2026

Pergeseran ini secara signifikan memperluas permukaan serangan (*attack surface*), mengingat agen AI kini beroperasi dengan tingkat hak akses yang eskalatif (Ying et al., 2026). Berpindahnya beban kerja dari sistem berbasis *cloud* yang terisolasi menuju agen otonom yang di-host secara lokal membawa risiko siber dan implikasi otonomi yang jauh lebih kompleks (Mathew et al., 2026).

Salah satu kerangka kerja agen otonom yang mendominasi ekosistem saat ini adalah OpenClaw. Sistem ini memberikan fleksibilitas luar biasa kepada pengguna dengan mengintegrasikan eksekusi shell, otomatisasi peramban, dan akses sistem file langsung ke dalam alur penalaran LLM (Zha et al., n.d.). Namun, akses istimewa yang menjadi keunggulan operasional ini secara inheren mengintroduksi kerentanan kritis (Deng et al., 2026). Hak akses yang tinggi dapat mengubah kesalahan interpretasi model menjadi ancaman tingkat sistem yang nyata, seperti kebocoran data sensitif, eskalasi hak istimewa, hingga eksekusi kode jarak jauh (*Remote Code Execution*). Laporan keamanan terbaru mengindikasikan bahwa sistem berarsitektur serupa OpenClaw sangat rentan terhadap injeksi *prompt*, di mana masukan yang tidak terpercaya (*untrusted input*) seperti pesan dari aplikasi obrolan dapat memanipulasi agen untuk melakukan tindakan destruktif di luar kendali pengguna (Wang et at., 2026).

Secara arsitektural, model keamanan bawaan OpenClaw didesain berdasarkan satu batas kepercayaan per gateway (*single trust boundary*). Artinya, sistem ini tidak dirancang untuk memfasilitasi banyak pengguna yang saling tidak percaya (*hostile multi-tenant*) (Suwansathit et al., 2026). Sayangnya, tingginya tingkat fleksibilitas *deployment* sering kali memicu implementasi konfigurasi yang berisiko, seperti eksposur jaringan yang terlalu terbuka, pengabaian *allowlist*, atau kegagalan dalam menerapkan prinsip hak istimewa minimal (*least privilege*). Laporan audit industri bahkan menyoroti fenomena kerentanan sistemik pada ekosistem OpenClaw, di mana kombinasi antara akses data lokal, penerimaan konten yang tidak terpercaya, dan eksekusi alat otomatis menciptakan risiko eksponensial(Audits & Reveal, 2026).

Berbagai literatur terkini telah merespons ancaman sistemik agen otonom ini melalui pendekatan teknis yang terfokus pada arsitektur pertahanan (*state of the art*). Sebagai contoh, penelitian Liu et al. (2026) berfokus pada mitigasi ancaman dengan mengusulkan ClawKeeper, sebuah lapisan perlindungan keamanan yang beroperasi secara real-time untuk mencegat eksekusi skill berbahaya. Di sisi lain, kajian oleh Liu et al. (2026) mencoba menambal kerentanan ini dari sudut pandang rekayasa perangkat lunak (*software engineering*) dengan mengusulkan arsitektur *defensible design.* Meskipun kontribusi kedua penelitian tersebut sangat krusial, terdapat celah analisis (*gap analysis*) yang fundamental.

kajian-kajian sebelumnya mengasumsikan solusi keamanan semata-mata sebagai perbaikan algoritma, modifikasi kode, atau penambahan modul filter sistem, sehingga secara keliru mereduksi bahkan mengabaikan variabel operasional manusia. Penelitian Li et al. (2026) maupun Wang et al. (2026) luput membedah realitas bahwa arsitektur sistem atau desain pertahanan yang diklaim "aman" pada tataran kode, tetap akan berujung pada eksploitasi fatal apabila praktisi IT secara sadar melonggarkan batasan otorisasi di lapangan. Oleh karena itu, kebaruan (*novelty*) dari penelitian ini terletak pada pergeseran fokus analitis yang radikal: dari mitigasi kerentanan mesin menuju akuntabilitas etis manusia. Berbeda dengan penelitian sebelumnya, studi ini menempatkan kedisiplinan dan keputusan etis seorang profesional IT saat melakukan *deployment* sebagai garda pertahanan utama yang tidak dapat disubstitusi oleh *patch* perangkat lunak mana pun.

Di titik inilah isu keamanan teknis bertransformasi menjadi dilema etika profesional yang mendesak. Kegagalan seorang praktisi IT dalam memahami batasan keamanan model bawaan misalnya memaksakan arsitektur *single-user* untuk melayani lingkungan multi-user bukanlah sekadar kelalaian teknis operasional (*technical glitch*), melainkan sebuah pelanggaran etika profesional. Pengoperasian sistem agen AI berhak akses tinggi menuntut kedisiplinan tingkat lanjut (Weidener et al., n.d.). Tanggung jawab keamanan tidak lagi dapat dibebankan sepenuhnya pada pengembang (*developer*) kerangka kerja, melainkan menjadi tanggung jawab moral dan profesional bersama bagi para pengguna yang melakukan *deployment*. Terlepas dari fitur *security* audit dan kontrol bawaan yang tersedia pada OpenClaw, kelalaian teknis yang berujung pada eksploitasi sistem akan berdampak langsung pada hilangnya privasi data, runtuhnya kerahasiaan, dan krisis akuntabilitas.

Berdasarkan celah penelitian dan urgensi tersebut, rumusan masalah utama dalam penelitian ini memfokuskan pada analisis tentang bagaimana OpenClaw menerapkan kontrol keamanan, bagaimana risiko muncul pada instalasi lokal, serta bagaimana risiko tersebut berdampak pada etika profesional dalam penggunaan agen AI. Untuk menjawab rumusan masalah tersebut, studi kasus kualitatif ini bertujuan mengevaluasi penerapan model keamanan bawaan dan efektivitas konfigurasi sistem (seperti kebijakan *loopback-only, token, dan allowlist*) pada lingkungan instalasi lokal. Selain itu, penelitian ini juga mengobservasi secara empiris manifestasi risiko keamanan saat sistem merespons *untrusted input*, guna menganalisis implikasi etika profesional terkait pembatasan akses, kerahasiaan data, dan distribusi letak tanggung jawab (*accountability*) antara pengembang dan praktisi IT (Chen et al., 2026).

2. KAJIAN TEORITIS

Transisi komputasi dari model bahasa besar (LLM) konvensional menuju agen AI otonom (*autonomous agents*) seperti OpenClaw telah secara fundamental mengubah landasan teori keamanan siber. Secara arsitektural, agen otonom tidak lagi beroperasi di ruang terisolasi, melainkan dibekali hak akses tingkat sistem yang memunculkan kerangka ancaman baru bernama *Lethal Trifecta* sebuah kombinasi eksponensial dan berisiko tinggi antara keleluasaan akses data lokal, penerimaan masukan yang tidak terpercaya (*untrusted input*), serta eksekusi alat secara otomatis. Teori *Zero-Trust Architecture* (ZTA) dan prinsip hak istimewa terkecil (*least privilege*) menjadi krusial dalam mengurai fenomena ini, mengingat model keamanan bawaan agen pada dasarnya didesain berdasarkan satu batas kepercayaan tunggal (*single trust boundary*). Transformasi ontologis ini mempertegas bahwa kerentanan pada agen AI berhak akses tinggi bukanlah sekadar cacat pemrosesan bahasa mesin, melainkan peleburan batas aman arsitektural yang berpotensi fatal apabila diterapkan pada infrastruktur tanpa isolasi yang memadai.

Berbagai literatur terdahulu telah membedah masifnya eskalasi kerentanan teknis pada ekosistem agen AI, seperti kampanye eksploitasi berskala besar ClawHavoc dan insiden runtuhnya basis data Moltbook yang mengekspos jutaan token autentikasi. Meskipun kajian sebelumnya seperti pendekatan *defensible design* oleh Li et al. (2026) dan perlindungan *runtime* oleh Liu et al. (2026) telah berupaya menambal kerentanan ini, studi-studi tersebut memiliki celah karena hanya mengasumsikan solusi keamanan sebatas pada perbaikan algoritma, sehingga secara keliru mereduksi variabel operasional manusia. Untuk menjembatani kesenjangan tersebut, penelitian ini menggunakan teori Etika Profesional Teknologi Informasi (TI) khususnya kewajiban uji tuntas (*due diligence*) dan akuntabilitas sebagai landasan pengikat. Berangkat dari konstruksi teoritis ini, penelitian ini dibangun di atas postulat bahwa sekuat apa pun penambalan kerentanan perangkat lunak pada agen otonom, risiko keamanan siber tingkat lanjut pada sistem tidak akan pernah dapat dieliminasi tanpa adanya penegakan batas tanggung jawab etis dan kedisiplinan dari praktisi IT saat melakukan konfigurasi dan pelimpahan hak akses di lapangan.

3. METODE PENELITIAN

Penelitian ini menggunakan metode kualitatif dengan pendekatan studi kasus tunggal (*single case study approach*). Desain ini dipilih karena rumusan masalah difokuskan pada pertanyaan eksplanatori yakni "bagaimana" dan "mengapa" terkait fenomena kontemporer risiko keamanan agen *Artificial Intelligence* (AI) otonom dalam konteks kehidupan nyata, di

mana batas antara operasional teknis dan batasan etika profesional tidak terlihat secara tegas (Audits & Reveal, 2026). Penelitian ini murni bersifat kualitatif dan tidak melibatkan survei, kuesioner, maupun pengumpulan data kuantitatif. Fokus analisis diarahkan secara mendalam pada manifestasi risiko keamanan arsitektural pada instalasi lokal OpenClaw dan implikasi etikanya bagi praktisi IT yang melakukan *deployment*.

Penelitian ini menggunakan metode kualitatif dengan pendekatan studi kasus tunggal (*single case study approach*). Desain ini dipilih karena rumusan masalah difokuskan pada pertanyaan eksplanatori yakni "bagaimana" dan "mengapa" terkait fenomena kontemporer risiko keamanan agen *Artificial Intelligence* (AI) otonom dalam konteks kehidupan nyata, di mana batas antara operasional teknis dan batasan etika profesional tidak terlihat secara tegas (Audits & Reveal, 2026). Penelitian ini murni bersifat kualitatif dan tidak melibatkan survei, kuesioner, maupun pengumpulan data kuantitatif. Fokus analisis diarahkan secara mendalam pada manifestasi risiko keamanan arsitektural pada instalasi lokal OpenClaw dan implikasi etikanya bagi praktisi IT yang melakukan *deployment*.

Data Primer

Diperoleh melalui observasi empiris dan eksperimen langsung (*hands-on experiment*) terhadap satu instalasi kerangka kerja OpenClaw. Sesuai dengan natur operasional sistem tersebut, seluruh interaksi, proses eksekusi, dan hasil observasi terekam murni dalam bentuk keluaran *Command Line Interface* (CLI). Bukti empiris yang diekstraksi meliputi: tangkapan layar eksekusi perintah di terminal, modifikasi parameter dalam file konfigurasi lokal (seperti struktur ~/.openclaw/), *output* tekstual dari fitur *security* audit bawaan sistem, serta log CLI yang mencatat respons agen saat memproses pesan eksternal (*untrusted input*) dan menjalankan aksi otomatis.

Data Sekunder

Diperoleh melalui dua jalur utama. Pertama, studi literatur sistematis mengenai taksonomi keamanan siber pada *Large Language Models* (LLM) dan kerangka kerja etika professional Kedua, analisis konten dari diskusi pengguna pada forum daring (memanfaatkan dokumen hasil analisis yang sumbernya dari forum online dan komunitas agen otonom seperti *Moltbook*) (Deng et al., 2026). Data dari forum online ini sangat krusial karena menyajikan temuan kerentanan di dunia nyata (*real-world exploits*)seperti kampanye ClawHavoc dan distribusi malicious skills serta diskursus organik komunitas mengenai letak tanggung jawab keamanan (Liu et al., 2026).

Pengumpulan data primer dilakukan di dalam lingkungan eksperimen yang terisolasi (*sandboxed environment*). Peneliti bertindak sebagai profesional IT yang menyimulasikan proses *deployment* agen OpenClaw melalui dua skenario pengujian utama:

a. Skenario Konfigurasi Aman (*Secure Setup*): Mengimplementasikan rekomendasi keamanan pengembang dengan membatasi *trust boundary*. Skenario ini mencakup pengaktifan kebijakan *loopback-only* (127.0.0.1), penggunaan mode autentikasi token, dan penerapan *allowlist* identitas secara ketat.
b. Skenario Konfigurasi Berisiko (*Insecure Setup*): Melonggarkan prinsip hak istimewa minimal (*least privilege*) dengan membuka eksposur jaringan (0.0.0.0) dan menonaktifkan autentikasi. Skenario ini dirancang untuk mengobservasi secara langsung kerentanan agen terhadap injeksi perintah (*prompt injection*) dari saluran aplikasi obrolan dan potensi kompromi terhadap privasi data lokal.

Seluruh peringatan sistem, perubahan status akses, dan jejak eksekusi didokumentasikan secara presisi dari antarmuka CLI selama eksperimen berlangsung. Bersamaan dengan itu, data sekunder dikumpulkan melalui ekstraksi dokumen hasil analisis diskursus dan laporan ancaman keamanan dari forum komunitas daring terkait.

Penelitian ini menerapkan teknik Analisis Konten Kualitatif (*Qualitative Content Analysis*) untuk memproses dan menginterpretasikan himpunan data teks yang didominasi oleh log CLI, skrip konfigurasi, dokumen audit, serta transkrip hasil analisis forum online (Ying et al., 2026). Data dianalisis secara tematik dan disintesis ke dalam lima area fokus utama, yaitu: (1) evaluasi model batasan aman (*single trust boundary*); (2) komparasi implementasi prinsip *least privilege*; (3) kerentanan penanganan untrusted input; (4) interpretasi peringatan *security* audit di CLI; dan (5) sintesis etika profesional terkait akuntabilitas antara pihak pengembang dan pengguna sistem.

4. HASIL DAN PEMBAHASAN

Hasil observasi dan simulasi pada infrastruktur lokal OpenClaw menunjukkan bahwa sistem agen kecerdasan buatan ini memberikan tingkat fleksibilitas operasional yang sangat tinggi, namun dengan batasan keamanan yang secara fundamental bergantung pada keputusan implementasi (*deployment*) dari profesional Teknologi Informasi (TI). Berdasarkan analisis terhadap konfigurasi jaringan dan autentikasi, model keamanan bawaan sistem ini dirancang dengan paradigma asisten personal yang mengandalkan batasan kepercayaan tunggal (*single trust boundary*). Temuan observasi pada status jaringan (*socket statistics*) memperlihatkan

bahwa secara asali (*default*), layanan mengikat pada alamat *loopback* (127.0.0.1) di porta 18789, yang mendedikasikan akses secara eksklusif untuk klien lokal. Namun, fleksibilitas konfigurasi memungkinkan profesional TI untuk mengubah ikatan antarmuka menjadi terbuka (0.0.0.0) tanpa persyaratan autentikasi tambahan. Perbandingan parameter konfigurasi teknis yang berdampak langsung pada postur keamanan sistem disajikan pada Tabel 1.

Tabel 1. Perbandingan parameter konfigurasi jaringan dan autentikasi pada instalasi OpenClaw.

Parameter Konfigurasi	Mode Instalasi Aman (Sesuai Prinsip)	Mode Instalasi Berisiko (Multi-pengguna Terbuka)	Implikasi Teknis
bind (Alamat Jaringan)	*loopback* (127.0.0.1)	0.0.0.0	Membuka visibilitas layanan dari *localhost* ke seluruh jaringan publik/lokal.
auth.mode (Autentikasi)	token	none	Menghilangkan validasi kredensial bagi entitas yang mengakses *endpoint* layanan.
Pengaturan Kontrol Antarmuka	Autentikasi berbasis identitas perangkat	gateway.controlUi.allowInsecureAuth=true	Melewati pembatasan akses kontrol *dashboard* pada konteks non-HTTPS.

Berdasarkan Tabel 1, terlihat jelas bahwa kontrol keamanan asali sebenarnya telah disediakan oleh pengembang, namun pemeliharaan integritas sistem sangat menuntut kedisiplinan dan kompetensi profesional TI. Mengubah arsitektur yang dirancang untuk pengguna tunggal menjadi layanan multi-pengguna yang saling tidak percaya (*hostile multi-tenant*) dengan membuka akses jaringan (0.0.0.0) dan mematikan fungsi token, merupakan pelanggaran langsung terhadap prinsip hak istimewa terkecil (*least privilege*). Dalam konteks etika profesional, tindakan modifikasi ini memindahkan tanggung jawab keamanan sepenuhnya ke tangan administrator. Ketika profesional TI mengabaikan pedoman instalasi aman demi kemudahan akses atau ketersediaan layanan (*availability*), mereka secara sadar menciptakan celah yang memungkinkan penyalahgunaan akses tingkat sistem oleh pihak yang tidak berkepentingan. Evaluasi terhadap literatur terdahulu juga sejalan dengan fenomena ini, di mana kelalaian konfigurasi pada lapisan gateway sering kali menjadi vektor utama dalam eksploitasi agen otonom (Audits & Reveal, 2026).

Lebih lanjut, risiko keamanan teknis ini dengan cepat bertransformasi menjadi risiko etika dan pelanggaran privasi ketika agen dihadapkan pada masukan tidak terpercaya (*untrusted input*) (Donge et al., 2026). Pada pengujian respons sistem terhadap instruksi

manipulatif, seperti permintaan untuk menampilkan seluruh file di sistem atau mengakses kredensial yang tersimpan, agen OpenClaw menunjukkan respons penolakan yang sesuai dengan pembatasan direktori kerjanya (/home/qudz/.openclaw/workspace). Sistem secara eksplisit menyatakan ketidakmampuannya mengakses informasi sensitif demi menjaga privasi pengguna. Namun, observasi pada struktur direktori utama (~/.openclaw/) mengungkap bahwa sistem ini memusatkan penyimpanan data yang sangat kritikal, termasuk direktori agen, integrasi pesan (Telegram), riwayat eksekusi, serta kredensial API. Meskipun agen memiliki pembatasan internal berbasis *prompt* dan pembatasan *sandbox* pada direktori kerjanya, isolasi ini dapat dengan mudah runtuh apabila parameter eksekusi diubah oleh administrator, misalnya dengan mengaktifkan akses *runtime* ke proses (*exec*) secara luas. Kegagalan sistem dalam memitigasi akses file ini bukan hanya kegagalan teknis, melainkan pelanggaran etis berat terkait kerahasiaan (*confidentiality*) dan akuntabilitas profesional TI dalam melindungi data penggunanya.

Kemampuan sistem untuk mendeteksi anomali dan memperingatkan pengguna menjadi fokus observasi berikutnya melalui pemanfaatan fitur audit keamanan terintegrasi. Analisis mendalam (*deep audit*) pada lingkungan simulasi menghasilkan peringatan yang secara langsung menyoroti celah dari konfigurasi buatan pengguna, yang secara rinci dijabarkan pada Tabel 2.

Tabel 2. Hasil audit keamanan fitur bawaan OpenClaw pada lingkungan simulasi.

Tingkat Keparahan	Modul Terdampak	Deskripsi Kerentanan dan Temuan Sistem
CRITICAL	channels.telegram	Tidak ada daftar putih (*allowlist*) pengirim pada grup. Anggota grup mana pun dapat memicu eksekusi perintah alat (*tool commands*).
WARN	gateway.trusted_proxies	Jaringan proksi terbalik (*reverse proxy*) tidak divalidasi, memungkinkan pemalsuan (*spoofing*) identitas klien lokal.
WARN	security.trust_model	Deteksi heuristik pengaturan multi-pengguna pada model asisten personal; *sandbox* dimatikan pada konteks agen secara asali.
WARN	config.insecure_flags	Parameter bendera berbahaya (*dangerous flags*) diaktifkan secara manual oleh pengguna (contoh: *insecure auth*).

Tabel 2 mempertegas argumen bahwa pengembang perangkat lunak (OpenClaw) telah mengimplementasikan mekanisme transparansi yang memadai terkait risiko infrastruktur. Peringatan tingkat kritis pada integrasi Telegram, misalnya, menunjukkan bahwa aktivasi bot tanpa pendefinisian daftar putih otorisasi (*allowlist*) secara langsung mengekspos fungsi agen

kepada publik. Selain itu, sistem secara cerdas mendeteksi adanya upaya memaksakan model asisten personal ke dalam skenario multi-pengguna, yang ditandai dengan peringatan pada modul *security*. *trust_model* Sistem audit merekomendasikan pemisahan batasan kepercayaan melalui gateway yang berbeda atau penerapan isolasi *sandbox* secara total jika sistem memang sengaja dibagikan (Sunil et al., 2026). Dalam lanskap etika keprofesian, temuan ini menyimpulkan adanya tanggung jawab yang ditanggung bersama (*shared responsibility*). Pengembang bertanggung jawab menyediakan alat pelindung (seperti *loopback*, token, dan peringatan audit), sementara profesional TI memikul kewajiban etis tak terbantahkan untuk merespons peringatan tersebut, mempertahankan arsitektur keamanan, dan tidak mengorbankan proteksi privasi demi fungsionalitas semata. Keengganan untuk mematuhi peringatan teknis dari hasil audit bukan sekadar kesalahan operasional Li et al. (2026) melainkan sebuah kelalaian profesional yang secara langsung membahayakan integritas data organisasi dan melanggar kode etik perlindungan informasi klien.

5. KESIMPULAN DAN SARAN

Secara umum, penelitian ini menyimpulkan bahwa integrasi agen kecerdasan buatan otonom berskala sistem seperti OpenClaw memberikan kapabilitas dan fleksibilitas operasional tingkat tinggi, namun secara bersamaan menghadirkan kerentanan sistemik yang sepenuhnya bergantung pada keputusan implementasi penggunanya. Celah keamanan yang muncul dalam ekosistem agen otonom *self-hosted* ini sejatinya bukanlah semata-mata kecacatan teknis dari desain perangkat lunak, melainkan merupakan manifestasi dari kelalaian etika profesional TI dalam menerapkan prinsip-prinsip keamanan siber yang fundamental. Tanggung jawab mitigasi risiko dan perlindungan privasi data telah melampaui batasan penyediaan fitur keamanan oleh pengembang, bergeser menjadi kewajiban moral dan keprofesian bagi para praktisi TI untuk mengonfigurasi, mengawasi, serta membatasi hak akses agen kecerdasan buatan secara ketat dan penuh kehati-hatian.

Menjawab tujuan-tujuan penelitian yang telah ditetapkan, temuan observasi menegaskan bahwa model keamanan asali OpenClaw dirancang secara eksklusif dengan batasan kepercayaan tunggal (*single trust boundary*). Pemaksaan arsitektur ini ke dalam skenario multi-pengguna melalui modifikasi konfigurasi jaringan menjadi terbuka (0.0.0.0) dan penonaktifan autentikasi berbasis token merupakan pelanggaran langsung terhadap pedoman implementasi aman dan prinsip hak istimewa terkecil (*least privilege*). Lebih lanjut, pengujian terhadap masukan tidak terpercaya (*untrusted input*) membuktikan bahwa meskipun agen memiliki kontrol internal untuk melindungi data sensitif, pertahanan ini dapat runtuh dan

menyebabkan pelanggaran privasi manakala profesional TI secara gegabah memberikan hak eksekusi tak terbatas yang mengekspos direktori operasional kritikal. Namun demikian, kehadiran fitur audit keamanan bawaan yang mampu mendeteksi secara presisi setiap anomali konfigurasi membuktikan bahwa pengembang telah memberikan transparansi risiko yang memadai. Hal ini mengukuhkan implikasi etis bahwa terdapat pembagian tanggung jawab yang jelas; pengembang memikul kewajiban untuk menyediakan instrumen peringatan dan kontrol pelindung, sementara profesional TI memegang tanggung jawab etis mutlak untuk merespons peringatan tersebut, menghindari kompromi keamanan demi sekadar kemudahan akses, dan menjamin akuntabilitas atas setiap aksi otonom yang dieksekusi oleh sistem di bawah kendalinya.

Berdasarkan konklusi dan batasan observasi dalam studi kasus kualitatif tunggal ini, terdapat beberapa arah pengembangan yang esensial untuk dieksplorasi pada masa mendatang. Gagasan penelitian selanjutnya diharapkan dapat memperluas cakupan analisis melalui tinjauan kuantitatif guna memetakan prevalensi instalasi agen kecerdasan buatan otonom di ruang publik yang terekspos akibat miskonfigurasi. Selain itu, kajian lanjutan sangat diperlukan untuk merumuskan kerangka kerja terstandardisasi atau Standar Operasional Prosedur (SOP) etika terapan yang secara spesifik mengatur pedoman *deployment* kecerdasan buatan berhak akses tinggi bagi tenaga profesional TI. Studi komparatif terhadap kerangka kerja agen otonom lainnya juga direkomendasikan guna membangun arsitektur pertahanan zero-trust yang lebih komprehensif dalam menavigasi masa depan interaksi antara manusia dan agen kecerdasan buatan.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri (UIN) Syarif Hidayatullah Jakarta atas dukungan akademik dan fasilitas yang diberikan selama proses penelitian ini. Apresiasi khusus penulis sampaikan kepada Ibu Evy Nurmiati S.Kom, M.MSI., selaku dosen pembimbing yang telah memberikan arahan teknis, motivasi, serta bimbingan yang sangat berharga dalam penyusunan manuskrip ini.

Penulis juga menyampaikan terima kasih kepada tim Asosiasi Pengelola Jurnal Indonesia atas kesempatan publikasi dan bantuan dalam proses evaluasi naskah. Ucapan terima kasih juga ditujukan kepada rekan-sejawat dan seluruh pihak yang telah membantu dalam proses pengumpulan data, analisis risiko, serta evaluasi kerentanan keamanan pada sistem agen OpenClaw selama masa observasi. Akhir kata, penulis menghargai dukungan moral dari

keluarga dan teman-teman sehingga penelitian mengenai etika dan keamanan sistem agen AI otonom ini dapat diselesaikan tepat waktu. Semoga penelitian ini bermanfaat bagi pengembangan tata kelola keamanan informasi dan standar perancangan agen open-source di masa depan.

DAFTAR REFERENSI

Audits, S., & Reveal, W. T. (2026). *41% of the most popular OpenClaw skills have security vulnerabilities*.

Chen, E., Guan, C., Elshafiey, A., Zhao, Z., Zekeri, J., Shaibu, A. E., ... Wu, C. J. (2026). *OpenClaw AI agents as informal learners at Moltbook: Characterizing an emergent learning community at scale*. *arXiv*. http://arxiv.org/abs/2602.18832

Deng, X., Zhang, Y., Wu, J., Bai, J., Yi, S., Zou, Z., ... Li, Q. (2026). *Taming OpenClaw: Security analysis and mitigation of autonomous LLM agent threats*. *arXiv*. http://arxiv.org/abs/2603.11619

Dong, B., Feng, H., & Wang, Q. (2026). *ClawDrain: Exploiting tool-calling chains for stealthy token exhaustion in OpenClaw agents*. *arXiv*. http://arxiv.org/abs/2603.00902

Li, Z., Li, W., & Li, X. (2026). *Defensible design for OpenClaw: Securing autonomous tool-invoking agents*. *arXiv*. http://arxiv.org/abs/2603.13151

Liu, S., Li, C., Wang, C., Hou, J., Chen, Z., Zhang, L., ... Wang, Z. (2026). *ClawKeeper: Comprehensive safety protection for OpenClaw agents through skills, plugins, and watchers*. *arXiv*. http://arxiv.org/abs/2603.24414

Mathew, A., College, B., & Virginia, W. (2026). *From conversation to command execution: A comparative threat modeling and risk analysis of OpenClaw and ChatGPT*. *ISRG Journal of Engineering and Technology, 5894*(I), 43–46. https://doi.org/10.5281/zenodo.18811875

Sunil, B. D., Sinha, I., Maheshwari, P., Todmal, S., Mallik, S., & Mishra, S. (2026). *Memory poisoning attack and defense on memory-based LLM agents*. *arXiv*. http://arxiv.org/abs/2601.05504

Suwansathit, S., Zhang, Y., & Gu, G. (2026). *A systematic taxonomy of security vulnerabilities in the OpenClaw AI agent framework*. *arXiv*. http://arxiv.org/abs/2603.27517

Wang, Y., Gao, H., Niu, Z., Liu, Z., Zhang, W., Wang, X., & Lian, S. (2026). *A systematic security evaluation of OpenClaw and its variants*. *arXiv*. http://arxiv.org/abs/2604.03131

Wang, Y., Xu, F., Lin, Z., He, G., Huang, Y., Gao, H., ... Liu, Z. (2026). *From assistant to double agent: Formalizing and benchmarking attacks on OpenClaw for personalized local AI agents*. *arXiv*. http://arxiv.org/abs/2602.08412

Wang, Z., Tu, H., Zhang, L., Chen, H., Wu, J., Liu, X., ... Xie, C. (2026). *Your agent, their asset: A real-world safety analysis of OpenClaw*. *arXiv*. http://arxiv.org/abs/2604.04759

Weidener, L., Lee, P., Karlsson, M., & Noessler, K. (n.d.). *From agent-only social networks to autonomous scientific*.

Ying, Z., Yang, X., Wu, S., Song, Y., Qu, Y., Li, H., ... Liu, X. (2026). *Uncovering security threats and architecting defenses in autonomous agents: A case study of OpenClaw*. *arXiv*. http://arxiv.org/abs/2603.12644

Zhan, Q., Liang, Z., Ying, Z., & Kang, D. (n.d.). *INJECAGENT: Benchmarking indirect prompt injections in tool-integrated large language model agents*. https://github.com/