Jicom : Journal Information & Computer
Volume 4, Nomor 1, Januari 2026, hlm. 38-46
DOI : 10.32493/jicomisc.v2i1.37666

P-ISSN : 3031-2779
E-ISSN : 3026-4197

Analisis Komprehensifi Evolusi Rekayasa Sosial: Konvergensi Agentic AI dan Deepfiake dalam Ekosistem Digital serta Strategi Mitigasi Berbasis Zero Trust

Mahbub Baihaqi, Mukhamad Kharismaula, Zulham Syafrawi
[1,2,3,4,5)]Program Studi S1, Jurusan Teknik Informatika, Universitas Siber Asia
Email: [*1]mahbubbaihaqi@gmail.com, [2]mkharismaula@gmail.com, [3]zulham@gmail.com

(Naskah masuk: *16 Januari 2026*, diterima untuk diterbitkan: *30 Januari 2026*)

Abstrak: *Laporan penelitian ini menyajikan tinjauan sistematis mengenai transformasi fundamental dalam lanskap serangan Social Engineering (Rekayasa Sosial), yang telah berevolusi dari teknik manipulasi psikologis konvensional berbasis teks menjadi operasi siber otonom yang digerakkan oleh Artificial Intelligence (AI). Memasuki periode strategis tahun 2026, ekosistem keamanan digital didominasi oleh konvergensi antara Generative AI, teknologi real-time Deepfake, dan kemunculan Agentic AI yang memiliki kapabilitas untuk melakukan pengintaian dan eksekusi serangan secara mandiri tanpa intervensi manusia. Studi ini secara mendalam menganalisis dua studi kasus representatif: insiden penipuan Deepfake CFO di Hong Kong yang mengakibatkan kerugian finansial masif sebesar US$25,6 juta, serta fenomena serangan Injection pada sistem verifikasi biometrik e-KYC perbankan di Asia Tenggara. Melalui pendekatan kualitatif dengan metode tinjauan literatur sistematis dan analisis studi kasus, penelitian ini mengidentifikasi bahwa metode pertahanan tradisional seperti verifikasi visual dan liveness detection pasif tidak lagi memadai. Penelitian ini mengusulkan adopsi kerangka kerja pertahanan adaptif PREDICT yang mengintegrasikan prinsip Zero Trust Architecture, deteksi liveness aktif multi-modal, dan harmonisasi kepatuhan terhadap regulasi UU PDP di Indonesia serta standar global seperti EU AI Act. Temuan ini memberikan kontribusi teoretis dan praktis bagi pengembangan strategi keamanan siber nasional di tengah eskalasi ancaman berbasis kecerdasan buatan..*

Kata Kunci: *Agentic AI, Deepfake, Keamanan Siber, Rekayasa Sosial, Zero Trust*

Abstract: *This research report presents a systematic review of the fundamental transformation in the landscape of Social Engineering attacks, which have evolved from conventional text-based psychological manipulation techniques to autonomous cyber operations driven by Artificial Intelligence (AI). Entering the strategic period of 2026, the digital security ecosystem is dominated by the convergence of Generative AI, real-time Deepfake technology, and the emergence of Agentic AI, which has the capability to conduct reconnaissance and execute attacks independently without human intervention. This study thoroughly analyzes two representative case studies: the Deepfake CFO fraud incident in Hong Kong, which resulted in massive financial losses of US$25.6 million, and the phenomenon of Injection attacks on the e-KYC biometric verification system of banks in Southeast Asia. Through a qualitative approach using systematic literature review and case study analysis, this research identifies that traditional defense methods such as visual verification and passive liveness detection are no longer adequate. This study proposes the adoption of the PREDICT adaptive defense framework, which integrates the principles of Zero Trust Architecture, multi-modal active liveness detection, and harmonization of compliance with Indonesia's PDP Law and global standards such as the EU AI Act. This finding contributes theoretically and practically to the development of national cybersecurity strategies amid escalating threats based on artificial intelligence.*

Keywords: *Agentic AI, Deepfake, Cybersecurity, Social Engineering, Zero Trust*

1. PENDAHULUAN

Dalam sejarah panjang keamanan informasi, elemen manusia secara konsisten diidentifikasi sebagai "rantai terlemah" (*the weakest link*) dalam arsitektur pertahanan siber organisasi. Serangan rekayasa sosial (*social engineering*), yang secara fundamental didefinisikan sebagai seni memanipulasi individu untuk mengungkapkan informasi rahasia atau melakukan tindakan yang mengkompromikan keamanan, telah

menjadi vektor serangan dominan selama beberapa dekade terakhir. Secara tradisional, serangan ini sangat bergantung pada interaksi manusia-ke-manusia, memanfaatkan kerentanan psikologis mendasar seperti kepercayaan, ketakutan, urgensi, dan altruisme atau keinginan untuk membantu. Namun, periode transisi antara tahun 2024 hingga 2026 menandai sebuah titik infleksi radikal dalam sejarah kejahatan siber, di mana paradigma serangan bergeser dari operasi manual yang padat karya menjadi operasi otomatis yang didukung oleh kecerdasan buatan (*Artificial Intelligence* - AI). Fenomena ini bukan sekadar peningkatan skala, melainkan perubahan fundamental dalam sifat ancaman itu sendiri, menciptakan apa yang disebut sebagai *Hyper-Social Engineering*.

Lanskap ancaman siber saat ini tidak lagi sekadar menghadapkan organisasi pada penipu manusia yang mengirimkan email *phishing* dengan tata bahasa yang buruk atau skema penipuan "Pangeran Nigeria" yang mudah diidentifikasi secara visual. Sebaliknya, organisasi dan individu kini berhadapan dengan entitas mesin cerdas yang mampu meniru identitas manusia dengan presisi yang hampir sempurna, mengaburkan batas ontologis antara realitas fisik dan fabrikasi digital. Data empiris mendukung urgensi fenomena ini; laporan ancaman global tahun 2025 mengindikasikan lonjakan signifikan dalam serangan siber di kawasan Asia Pasifik, termasuk Indonesia. Badan Siber dan Sandi Negara (BSSN) mencatat peningkatan anomali trafik yang menyasar sektor vital seperti perbankan, energi, dan pemerintahan, dengan perkiraan bahwa 60% dari serangan tersebut memanfaatkan teknologi AI untuk menembus pertahanan konvensional. Hal ini menunjukkan bahwa AI telah mendemokratisasi kemampuan serangan canggih, memungkinkan aktor ancaman dengan keterampilan teknis rendah untuk melancarkan serangan setingkat *nation-state*.

Transisi dari serangan manual ke serangan otomatis berbasis AI telah mengubah fundamental ekonomi kejahatan siber secara drastis (*cybercrime economics*). Di masa lalu, pelaku ancaman harus menginvestasikan waktu berjam-jam untuk melakukan pengintaian (*reconnaissance*) dan menyusun pesan yang dipersonalisasi untuk target bernilai tinggi (*spear-phishing*). Proses ini mahal dan tidak dapat diskalakan dengan mudah. Kini, dengan kemunculan *Large Language Models* (LLM) dan *Generative AI* (GenAI), pelaku dapat menghasilkan ribuan konten penipuan yang sangat personal (*hyper-personalized*) dalam hitungan detik dengan biaya marjinal mendekati nol. AI mampu memproses jejak digital korban dari berbagai platform media sosial, menganalisis gaya komunikasi, preferensi, dan jaringan profesional mereka, kemudian menyusun narasi penipuan yang secara kontekstual sangat relevan dan sulit dibedakan dari komunikasi yang sah.

Lebih jauh lagi, evolusi teknologi *Deepfake* audio dan video telah meruntuhkan pilar kepercayaan paling mendasar dalam komunikasi digital: prinsip "melihat adalah percaya" (*seeing is believing*) tidak lagi valid. Insiden di mana seorang karyawan mentransfer dana jutaan dolar setelah melakukan panggilan video dengan "atasannya" yang ternyata adalah proyeksi AI menunjukkan tingkat ancaman yang belum pernah terjadi sebelumnya. Teknologi ini mengeksploitasi bias kognitif manusia yang secara evolusioner terprogram untuk mempercayai input visual dan auditori langsung. Ketika indra penglihatan dan pendengaran dimanipulasi dengan fidelitas tinggi, mekanisme pertahanan logis manusia sering kali gagal berfungsi, terutama di bawah tekanan urgensi yang direkayasa oleh penyerang.

Penelitian ini difokuskan untuk menjawab kompleksitas ancaman yang terus berevolusi ini melalui tiga rumusan masalah utama yang mendesak. Pertama, bagaimana evolusi spesifik teknologi *Generative AI* dan *Deepfake* mengubah taktik serangan rekayasa sosial secara teknis dan operasional pada periode 2024-2026? Kita perlu memahami transformasi arsitektur serangan yang memungkinkan manipulasi menjadi lebih canggih dan sulit dideteksi. Kedua, apa dampak nyata dan terukur dari serangan *Agentic AI* serta manipulasi biometrik terhadap sektor korporat dan finansial, khususnya di kawasan Asia Tenggara? Melalui studi kasus konkret, penelitian ini akan mengkuantifikasi kerugian finansial dan reputasi yang ditimbulkan. Ketiga, sejauh mana efektivitas kerangka hukum yang ada, seperti UU ITE dan UU PDP di Indonesia serta EU AI Act secara global, dalam memitigasi ancaman ini? Evaluasi kritis terhadap instrumen hukum diperlukan untuk menentukan apakah regulasi saat ini mampu mengejar kecepatan inovasi kejahatan siber.

Tujuan dari artikel ini adalah mendokumentasikan transformasi lanskap ancaman secara sistematis, menganalisis mekanisme teknis dan psikologis di balik serangan modern, serta merumuskan strategi pertahanan komprehensif yang menggabungkan aspek teknis, psikologis, dan regulasi hukum. Penelitian ini diharapkan dapat memberikan panduan strategis bagi praktisi keamanan siber, pengambil kebijakan, dan akademisi dalam membangun ekosistem digital yang lebih tangguh (resilien) menghadapi ancaman hibrida masa depan.

2. METODE PENELITIAN

Untuk menjawab permasalahan yang kompleks dan multidimensi mengenai evolusi rekayasa sosial berbasis AI, penelitian ini mengadopsi pendekatan kualitatif dengan desain deskriptif-analitis. Metodologi ini dipilih karena kemampuannya untuk mengeksplorasi fenomena yang sedang berkembang (*emerging phenomena*) di mana variabel-variabelnya belum sepenuhnya terdefinisi dan terus berubah secara dinamis. Kerangka kerja metodologis terdiri dari tiga komponen utama: Tinjauan Literatur Sistematis (*Systematic Literature Review*), Analisis Studi Kasus (*Case Study Analysis*), dan Analisis Komparatif Regulasi (*Comparative Regulatory Analysis*).

2.1. Tinjauan Literatur Sistematis (Systematic Literature Review):

1. Strategi Pencarian: Pencarian literatur dilakukan pada basis data akademik terkemuka termasuk IEEE Xplore, ACM Digital Library, ScienceDirect, dan Google Scholar. Kata kunci yang digunakan meliputi kombinasi dari "Social Engineering", "Generative AI", "Deepfake", "Agentic AI", "Cybersecurity", "Zero Trust", dan "Biometric Injection Aflack"

2. Kritaria Inklusi dan Eksklusi
   - *Inklusi:* Artikel jurnal peer-review, prosiding konferensi internasional (IEEE, ACM), laporan teknis dari vendor keamanan siber terkemuka (CrowdStrike, Palo Alto Networks, Group-IB), dan dokumen regulasi resmi yang diterbitkan antara tahun 2022 hingga 2026. Fokus materi harus pada konvergensi AI dan keamanan siber.
   - *Eksklusi:* Artikel yang diterbitkan sebelum tahun 2022 (kecuali teori dasar psikologi), artikel yang tidak berbahasa Inggris atau Indonesia, dan publikasi non-teknis yang tidak memiliki landasan data yang kuat

3. Ekstraksi Data: Data diekstraksi berdasarkan kategori: evolusi vektor serangan, mekanisme teknis *Deepfake*, dampak kerugian finansial, dan efektivitas metode mitigasi yang ada

2.2. Studi Kasus (Case Study Analysis)

Penelitian ini menggunakan metode studi kasus ganda (*multiple case study*) untuk menganalisis insiden keamanan siber nyata yang merepresentasikan tren global dan regional. Dua kasus utama dipilih berdasarkan signifikansi dampak dan kebaruan modus operandi:

1. **Studi Kasus Global:** Insiden penipuan *Deepfake* CFO di Hong Kong tahun 2024.Analisis difokuskan pada rekonstruksi kronologi serangan, identifikasi titik kegagalan manusia dan teknologi, serta analisis teknik *social engineering* tingkat lanjut yang digunakan pelaku.
2. **Studi Kasus Regional:** Fenomena serangan *Injection* pada sistem e-KYC perbankan di Asia Tenggara dan Indonesia. Analisis difokuskan pada aspek teknis manipulasi biometrik, kelemahan arsitektur aplikasi *mobile banking*, dan implikasi terhadap kepercayaan ekosistem keuangan digital

2.3. Analisis Komparatifi Regulasi

Analisis ini membandingkan kerangka hukum di Indonesia dengan standar internasional untuk mengidentifikasi kesenjangan kebijakan. Instrumen hukum utama yang dianalisis meliputi:

- **Indonesia:** Undang-Undang Informasi dan Transaksi Elektronik (UU ITE) Nomor 1 Tahun 2024 dan Undang-Undang Perlindungan Data Pribadi (UU PDP) Nomor 27 Tahun 2022.
- Global: EU AI Act (Uni Eropa) sebagai tolok ukur regulasi AI komprehensif pertama di dunia.

Analisis difokuskan pada pasal-pasal yang mengatur manipulasi konten sintetis, perlindungan data biometrik, dan sanksi bagi penyalahgunaan AI.1

2.4. Instrumen dan Validasi Data

Instrumen utama dalam penelitian ini adalah peneliti sendiri (*human instrument*) yang dibantu oleh matriks ekstraksi data. Validitas data dipastikan melalui triangulasi sumber, yaitu membandingkan data dari laporan industri, literatur akademik, dan pemberitaan media terverifikasi untuk mendapatkan pemahaman yang

utuh dan objektif mengenai fenomena yang diteliti. Analisis data dilakukan secara induktif, bergerak dari temuan spesifik menuju generalisasi pola serangan dan rekomendasi strategis.

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan temuan komprehensif mengenai transformasi teknis serangan rekayasa sosial, analisis mendalam terhadap insiden dunia nyata, dekonstruksi psikologis korban, serta evaluasi kerangka hukum yang berlaku

3.1. Evolusi Teknis: Dari Phishing Statis ke Otonomi Agentic AI

Transformasi serangan siber tidak terjadi dalam ruang hampa, melainkan mengikuti kurva perkembangan teknologi AI. Analisis kami mengidentifikasi tiga fase evolusi utama yang mengubah lanskap ancaman secara fundamental.

3.1.1. Era Generative AI: Personalisasi Skala Besar

Fase pertama evolusi ditandai dengan integrasi *Generative AI* ke dalam siklus serangan *phishing*. Pada era pra-AI, serangan *phishing* sering kali mudah dikenali karena karakteristiknya yang generik, kesalahan linguistik yang mencolok, dan visual yang tidak profesional. Hambatan ini membatasi efektivitas serangan hanya pada populasi yang kurang terliterasi secara digital. Namun, kehadiran *Large Language Models* (LLM) telah mengubah paradigma ini secara drastis.

Data penelitian tahun 2024-2025 menunjukkan lonjakan efektivitas yang mengkhawatirkan: email *phishing* yang dihasilkan oleh AI mencatat tingkat keberhasilan klik (*click-through rate*) sebesar 54%, jauh melampaui rata-rata 12% pada *phishing* manual. Kenaikan efektivitas ini didorong oleh kemampuan *Hyper-Personalization* yang dimiliki AI.

- **Pengumpulan Data Otomatis:** Algoritma AI mampu melakukan *scraping* data secara masif dari platform seperti LinkedIn, Twifler, dan Instagram untuk membangun profil target yang komprehensif.
- **Analisis Konteks dan Sentimen:** AI menganalisis konteks bisnis terkini seperti berita merger, peluncuran produk, atau perubahan manajemen dan menyisipkannya ke dalam narasi penipuan, menciptakan ilusi urgensi dan relevansi yang tinggi.
- **Peniruan Gaya Bahasa (*Style Mimicry*):** LLM dapat dilatih menggunakan sampel email atau tulisan target sebelumnya untuk meniru sintaksis, pilihan kata, dan nada bicara atasan atau rekan kerja, membuat deteksi berbasis anomali gaya bahasa menjadi tidak efektif.

3.1.2. Revolusi Deepfiake: Realitas Sintetis Waktu Nyata

Fase kedua melibatkan manipulasi modalitas audio dan visual. Teknologi *Deepfake* telah bergerak dari ranah pasca-produksi yang membutuhkan daya komputasi besar menjadi aplikasi waktu nyata (*real-time*) yang dapat dijalankan pada perangkat konsumen.

Tabel 1. Perbandingan Karakteristik Serangan Phishing Tradisional vs AI-Enhanced

Fitur Serangan	**Phishing Tradisional**	**AI-Enhanced Phishing / Deepfake**
Kualitas Bahasa	Sering terdapat kesalahan, kaku	Sempurna, natural, kontekstual
Personalisasi	Rendah (Generik/Template)	Tinggi (Hyper-Personalized)
Biaya Pembuatan	Rendah untuk massal, Tinggi untuk *Spear*	Sangat Rendah (Otomatisasi Penuh)
Media	Teks (Email, SMS)	Multi-modal (Teks, Suara, Video Real-time)

Target Psikologis	Ketakutan dasar, Keserakahan	Kepercayaan Visual, Otoritas, Hubungan Personal
Waktu Persiapan	Jam hingga Hari	Detik hingga Menit

Teknologi kloning suara seperti VALL-E 2 dari Microsoft mendemonstrasikan kemampuan *Zero-Shot Voice Cloning*, di mana model AI hanya membutuhkan sampel audio 3 detik untuk mensintesis suara target dengan intonasi dan emosi yang akurat. Ini memungkinkan serangan *Vishing* (*Voice Phishing*) yang sangat meyakinkan, di mana penyerang dapat meniru suara CEO atau anggota keluarga dalam situasi darurat. Lebih lanjut, *Deepfake* video *real-time* kini memungkinkan manipulasi wajah (*face swap*) dan sinkronisasi bibir (*lip-sync*) secara langsung dalam panggilan video, menghilangkan latensi yang sebelumnya menjadi indikator utama kepalsuan.

3.1.3. Agentic AI: Serangan Siber Otonom

Fase ketiga dan yang paling mutakhir adalah kemunculan *Agentic AI*. Berbeda dengan *Generative AI* yang bersifat pasif (menunggu *prompt*), *Agentic AI* memiliki agensi untuk bertindak secara mandiri mencapai tujuan yang ditetapkan. Dalam konteks ofensif, *Agentic AI* mampu mengotomatisasi seluruh rantai serangan (*kill chain*):

1. Pengintaian Otonom: Agen AI memindai permukaan serangan organisasi, mengidentifikasi karyawan rentan, dan memetakan arsitektur jaringan tanpa arahan manusia.
2. Perencanaan Adaptif: Agen menyusun strategi serangan bertahap. Jika satu vector (misalnya email) gagal, agen secara otomatis beralih ke vektor alternatif (misalnya pesan LinkedIn atau WhatsApp) berdasarkan pembelajaran *real-time*

Eksekusi Cepat: Kecepatan serangan meningkat eksponensial. Waktu rata-rata untuk eksfiltrasi data (*Mean Time to Exfiltrate*) menyusut dari hitungan hari menjadi menit, memberikan tekanan luar biasa pada tim *Security Operations Center* (SOC) yang masih mengandalkan respons manual.

3.2. Analisis Studi Kasus: Dampak pada Ekosistem Keuangan

Dua studi kasus berikut mengilustrasikan bagaimana teori ancaman di atas bermanifestasi menjadi kerugian nyata

3.2.1. Kasus Penipuan Deepfiake CFO di Hong Kong

Insiden yang menimpa perusahaan multinasional Arup di Hong Kong pada awal 2024 menjadi *wake-up call* global. Kerugian sebesar HK$200 juta (US$25,6 juta) terjadi akibat satu sesi konferensi video.

- Analisis Kejadian: Karyawan bagian keuangan menerima pesan dari "CFO" terkait transaksi rahasia. Meski awalnya curiga, keraguan karyawan tersebut sirna setelah mengikuti panggilan video grup. Dalam panggilan tersebut, hadir CFO dan beberapa eksekutif lain yang wajah dan suaranya sangat identik dengan aslinya.
- Faktor Kegagalan: Investigasi mengungkapkan bahwa semua peserta lain dalam panggilan video tersebut kecuali korban adalah *Deepfake* hasil rekayasa AI. Pelaku menggunakan rekaman video publik untuk melatih model wajah dan suara. Faktor krusial di sini adalah *Social Proof*; kehadiran banyak "saksi" (eksekutif lain) memvalidasi situasi palsu tersebut di mata korban.
- Implikasi: Kasus ini membuktikan bahwa verifikasi visual, yang selama ini dianggap sebagai standar emas autentikasi jarak jauh, telah dikompromikan sepenuhnya. Protokol keamanan yang hanya mengandalkan "mengenali wajah" tidak lagi relevan

3.2.2. Serangan Injection pada Sistem e-KYC di Asia Tenggara

Di kawasan Asia Tenggara, ancaman mengambil bentuk yang lebih teknis melalui serangan terhadap sistem *Electronic Know Your Customer* (e-KYC).

- **M**ekanisme Serangan: Pelaku tidak lagi menggunakan metode *Presentation AWack* (memperlihatkan foto/video ke kamera). Sebaliknya, mereka menggunakan Teknik *Injection AWack*. Dengan menggunakan perangkat yang telah di-*root* atau emulator, pelaku memasang *virtual camera driver* yang memanipulasi aliran data video pada level sistem operasi. Aplikasi perbankan "berpikir" sedang mengakses kamera fisik, padahal menerima aliran data video *Deepfake* yang telah disiapkan sebelumnya.
- Dampak Sistemik: Serangan ini memungkinkan pembuatan ribuan akun bank palsu (*mule accounts*) secara otomatis yang kemudian digunakan untuk pencucian uang dan pendanaan aktivitas ilegal. Metode deteksi *liveness* pasif (analisis tekstur kulit, kedipan mata) gagal mendeteksi serangan ini karena video *Deepfake* yang diinjeksi memang memiliki karakteristik visual manusia hidup. Ini menuntut perubahan mendasar dari deteksi pasif ke deteksi aktif dan verifikasi integritas perangkat keras (*hardware aWestation*)

3.3. Dekonstruksi Psikologis: Mengapa Pertahanan Manusia Gagal?

Keberhasilan serangan AI tidak hanya bergantung pada kecanggihan teknologi, tetapi juga pada eksploitasi arsitektur kognitif manusia. *Dual-Process Theory* yang dikemukakan oleh Kahneman dan Tversky memberikan kerangka analisis yang relevan.

1. Dominasi Sistem 1: Otak manusia beroperasi pada dua sistem: Sistem 1 (cepat, intuitif, emosional) dan Sistem 2 (lambat, analitis, logis). Serangan rekayasa sosial dirancang khusus untuk membajak Sistem 1. Dengan menciptakan skenario urgensi tinggi ("Transfer sekarang atau operasi berhenti!") atau ketakutan, penyerang memaksa otak korban memintas analisis kritis Sistem 2 dan langsung bereaksi menggunakan Sistem 1.
2. Visual Truth Bias: Evolusi ribuan tahun telah melatih otak manusia untuk mempercayai input visual sebagai kebenaran objektif. "Melihat" seseorang secara fisik adalah bukti keberadaan mereka. Teknologi *Deepfake* meretas bias kognitif purba ini. Ketika korban melihat wajah atasan mereka di layar, Sistem 1 secara otomatis memverifikasi identitas tersebut sebagai "asli", menciptakan *cognitive dissonance* yang sulit ditembus oleh keraguan logis.
3. Otoritas dan Kepatuhan: Prinsip persuasi Cialdini tentang Otoritas memainkan peran sentral. Dalam budaya korporat yang hierarkis, perintah dari C-level Executive (CEO/CFO) membawa bobot psikologis berat. Menolak atau memverifikasi ulang perintah atasan sering dianggap sebagai tindakan insubordinasi. *Deepfake* memperkuat tekanan ini dengan menghadirkan figur otoritas secara visual, membuat kepatuhan menjadi respons default.

3.4. Analisis Kerangka Hukum dan Regulasi

Respons hukum terhadap ancaman ini masih dalam tahap perkembangan, dengan variasi signifikan antara yurisdiksi nasional dan global

3.4.1. Regulasi di Indonesia (UU ITE dan UU PDP)

Indonesia memiliki instrumen hukum yang dapat diterapkan, meskipun belum secara eksplisit mengatur *Deepfake* secara komprehensif.

- UU ITE (Revisi 2024): Pasal 35 UU ITE tentang manipulasi data elektronik menjadi landasan utama. Pasal ini melarang penciptaan atau manipulasi informasi elektronik agar seolah-olah data yang otentik. Secara teoritis, pembuatan *Deepfake* untuk penipuan masuk dalam delik ini. Namun, tantangan utamanya adalah pembuktian forensik digital dan atribusi pelaku yang sering berada di yurisdiksi asing.

- UU PDP (2022): Memberikan perlindungan terhadap data biometrik (wajah, suara) sebagai data pribadi spesifik. Penggunaan wajah seseorang untuk melatih model AI tanpa *consent* merupakan pelanggaran. UU PDP memperkenalkan sanksi denda administratif hingga 2% dari pendapatan tahunan, yang bisa menjadi deterens bagi korporasi yang lalai, namun kurang efektif untuk sindikat kriminal anonym.

3.4.2. Standar Global (EU AI Act)

Sebagai pembanding, *EU AI Act* yang diadopsi Uni Eropa menetapkan standar yang lebih preskriptif. Pasal 50 mewajibkan transparansi penuh untuk sistem AI yang menghasilkan konten sintetis. Setiap konten *Deepfake* harus diberi label yang dapat dibaca mesin (*machine-readable watermarking*) dan dapat dideteksi sebagai artifisial. Pendekatan berbasis risiko ini kemungkinan akan menciptakan "Brussels Effect", di mana perusahaan teknologi global akan mengadopsi standar ini di seluruh dunia, termasuk Indonesia, untuk menyederhanakan kepatuhan

4. REKOMENDASI DAN STRATEGI MITIGASI

Menghadapi konvergensi ancaman ini, paradigma pertahanan "perimeter-centric" sudah usang. Organisasi memerlukan strategi pertahanan berlapis (*Defense in Depth*) yang adaptif. Kami mengusulkan kerangka kerja **PREDICT** (*Prevention, Response, Detection, Incident management, Continuous improvement, Training*) dengan penekanan pada aspek-aspek berikut:

4.1. Arsitektur Zero Trust dan Verifikasi Out-ofi-Band

Prinsip *Zero Trust* ("Never Trust, Always Verify") harus diimplementasikan secara radikal pada setiap interaksi digital.

- **Verifikasi Multi-Kanal (Out-of-Band):** Organisasi harus mewajibkan protokol verifikasi silang untuk setiap transaksi bernilai tinggi. Jika instruksi diterima melalui *video call*, verifikasi harus dilakukan melalui saluran terpisah yang aman, seperti telepon seluler terenkripsi atau aplikasi pesan internal, untuk memastikan identitas pengirim. "Melihat" tidak lagi cukup; konfirmasi teknis diperlukan.
- **Phishing-Resistant MFA:** Meninggalkan metode autentikasi berbasis SMS OTP atau aplikasi *authenticator* standar yang rentan terhadap *phishing*. Beralih ke kunci keamanan perangkat keras (FIDO2/WebAuthn) atau biometrik yang terikat perangkat (*device-bound passkeys*) yang secara kriptografis tidak dapat dikloning atau dipancing oleh situs palsu.

4.2. Evolusi Deteksi Biometrik: Melawan Injection

Untuk sektor perbankan dan layanan yang mengandalkan e-KYC, pembaruan teknologi deteksi mutlak diperlukan.

- **Active Liveness & Challenge-Response:** Mengganti deteksi pasif dengan tantangan aktif yang acak (*randomized challenge-response*). Contohnya, meminta pengguna memantulkan cahaya warna tertentu dari layar ponsel ke wajah (*flash challenge*) atau melakukan gestur mikro yang sangat spesifik dalam jendela waktu sempit. Hal ini mempersulit penggunaan *Deepfake* pra-rekaman.
- **Hardware Attestation:** Aplikasi harus mampu memverifikasi integritas perangkat keras sebelum menerima input kamera. Ini mencakup deteksi *rooting*, *jailbreak*, dan keberadaan *hooking framework* (seperti Frida) yang biasa digunakan untuk injeksi video. Analisis metadata sensor (akselerometer, giroskop) juga harus dikorelasikan dengan pergerakan video untuk memvalidasi bahwa video berasal dari sensor fisik

4.3. Ketahanan Manusia (Human Firewall Resilience)

Teknologi pertahanan terbaik akan gagal jika manusia di dalamnya tidak siap. Pelatihan kesadaran keamanan harus berevolusi dari sekadar mengenali *phishing* teks menjadi simulasi serangan AI.

- **Simulasi Serangan Deepfake:** Melakukan latihan rutin di mana karyawan "diserang" dengan simulasi *vishing* atau *deepfake* video untuk melatih respons mereka terhadap tekanan psikologis dan mengenali artefak visual halus (seperti *glitching* pada area telinga atau sinkronisasi bibir yang tidak natural).
- **Budaya Keamanan Psikologis:** Membangun budaya organisasi di mana memverifikasi instruksi atasan bahkan CEO diapresiasi sebagai tindakan keamanan proaktif, bukan pembangkangan. Ini menetralisir eksploitasi prinsip otoritas yang digunakan penyerang

5. KESIMPULAN

Penelitian ini menyimpulkan bahwa konvergensi antara *Generative AI*, *Deepfake*, dan *Agentic AI* telah menciptakan paradigma ancaman baru yang bersifat asimetris dan sangat skalabel. Kasus kerugian finansial di Hong Kong dan serangan injeksi massal di Asia Tenggara adalah manifestasi awal dari tren yang akan semakin intensif pada tahun 2026 dan seterusnya. Batas antara realitas dan rekayasa digital telah kabur, menuntut redefinisi konsep "kepercayaan" dalam interaksi siber.

Analisis membuktikan bahwa pertahanan tunggal baik itu teknologi biometrik pasif maupun pelatihan kesadaran konvensional tidak lagi memadai. Strategi mitigasi yang efektif menuntut pendekatan holistik yang mengintegrasikan arsitektur *Zero Trust* yang ketat, teknologi deteksi *liveness* aktif berbasis perangkat keras, dan kerangka hukum yang responsif terhadap dinamika teknologi. Bagi Indonesia, harmonisasi regulasi UU ITE dan UU PDP dengan standar teknis global serta peningkatan literasi digital nasional menjadi prasyarat mutlak untuk menjaga kedaulatan dan keamanan ekonomi digital di era kecerdasan buatan. Masa depan

keamanan siber tidak lagi tentang membangun tembok yang lebih tinggi, melainkan tentang membangun sistem dan manusia yang mampu memverifikasi kebenaran di tengah lautan manipulasi digital.

DAFTAR PUSTAKA

[1] **Al-Aswadi, F. N., et al. (2025).** Evolving Zero Trust architectures for AI-driven cyber threats in healthcare and other high-risk data environments: A systematic review. *International Journal of Environmental Research and Public Health*. https://pmc.ncbi.nlm.nih.gov/articles/PMC12229833/

[2] **Di Mauro, M., Casola, V., Choo, K. K. R., & Galdi, C. (2025).** The erosion of cybersecurity zero-trust principles through generative AI: A survey on the challenges and future directions. *Future Internet*, *17*(1), 5. https://doi.org/10.3390/fi17010005

[3] **Jaleel, A., et al. (2025).** Cognitive firewalls: Mitigating LLM-powered social engineering through personality-aware behavioral analytics. *The American Journal of Engineering and Technology*, *7*(01), 1-15. https://www.theamericanjournals.com/index.php/tajet/article/view/6943

[4] **Kaloudi, N., & Li, J. (2020).** The AI-driven threat landscape: A survey of dangers and management strategies. *ACM Computing Surveys*, *53*(1), 1-35. https://doi.org/10.1145/3372819

[5] **Nguyen, H. S., et al. (2025).** *EdgeDoc: Hybrid CNN-transformer model for accurate forgery detection and localization in ID documents*. arXiv. https://arxiv.org/abs/2508.16284

[6] **NIST. (2024).** *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-2

[7] **Shreshtha, S. (2025).** *Zero Trust architecture in AI-driven cybersecurity: A machine learning perspective*. ResearchGate. https://www.researchgate.net/publication/388523876_Zero_Trust_Architecture_in_AI-Driven_Cybersecurity_A_Machine_Learning_Perspective

[8] **Nguyen, H. S., et al. (2025).** *EdgeDoc: Hybrid CNN-transformer model for accurate forgery detection and localization in ID documents*. arXiv. https://arxiv.org/abs/2508.16284

[9] **Respati, A. A. (2024).** Reformulasi UU ITE terhadap Artificial Intelligence dibandingkan dengan Uni Eropa dan China AI Act regulation. *Jurnal USM Law Review*, *7*(3), 1737-1758. https://doi.org/10.26623/julr.v7i3.10578

[10] **Roy, S., et al. (2024).** Zero trust and AI: A synergistic approach to next-generation cyber threat mitigation. *World Journal of Advanced Research and Reviews*, *24*(03), 3374-3387. https://doi.org/10.30574/wjarr.2024.24.3.3883

[11] **World Economic Forum. (2026).** *Unmasking cybercrime: Strengthening digital identity verification against deepfakes*. https://www.weforum.org/reports/unmasking-cybercrime-strengthening-digital-identity-verification-against-deepfakes-2026

[12] **Zhang, X., et al. (2025).** *Zero Trust Architecture: A systematic literature review*. arXiv. https://doi.org/10.48550/arXiv.2503.11659