

Analisis Ancaman Social Engineering Berbasis Agentic AI dan Strategi Mitigasi Menggunakan Pendekatan Zero Trust

Teguh Pradana^{1*}, Nur Amalia Hasma², Handry Eldo³, Renita Selviana⁴, Ayuliana⁵

^{1*} Program Studi Sistem Dan Teknologi Informasi, Fakultas Teknik, Institut Teknologi dan Bisnis Yadika Pasuruan, Kabupaten Pasuruan, Provinsi Jawa Timur, Indonesia.

² Program Studi Informatika, Fakultas Komputer dan Multimedia, Universitas Islam Kebangsaan Indonesia, Kabupaten Bireuen, Provinsi Aceh, Indonesia.

³ Program Studi Teknologi Informasi, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah mahakarya Aceh, Kabupaten Bireuen, Provinsi Aceh, Indonesia.

⁴ Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Institut Teknologi dan Bisnis Yadika Pasuruan, Kabupaten Pasuruan, Jawa Timur, Indonesia.

⁵ Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Bina Nusantara, Kota Jakarta Barat, Jakarta Barat, Indonesia.

Corresponding Email: teguh_p@stmik-yadika.ac.id^{1*}, nuramaliahasma05@gmail.com², handry.eldo@gmail.com³, renita.selvi@itbyadika.ac.id⁴, ayuliana_st@binus.ac.id⁵

Histori Artikel:

Dikirim 26 Agustus 2026; Diterima dalam bentuk revisi 15 September 2026; Diterima 25 September 2026; Diterbitkan 28 September 2025. Semua hak dilindungi oleh Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) STMik Indonesia Banda Aceh.

Abstrak

Perkembangan *Agentic Artificial Intelligence (Agentic AI)* menghadirkan tantangan baru dalam keamanan siber karena kemampuannya dalam melakukan perencanaan, adaptasi, dan tindakan secara otonom. Penelitian ini bertujuan menganalisis ancaman *social engineering* berbasis *Agentic AI* serta merumuskan strategi mitigasinya menggunakan pendekatan *Zero Trust*. Metode penelitian yang digunakan adalah studi literatur dan pemodelan ancaman untuk mengidentifikasi vektor serangan, kerentanan pengguna, serta potensi dampaknya. Hasil analisis menunjukkan bahwa *Agentic AI* dapat meningkatkan personalisasi, skala, kecepatan, dan efektivitas serangan *social engineering*. Penelitian ini menghasilkan kerangka konseptual mitigasi *social engineering* berbasis *Agentic AI* dengan prinsip *Zero Trust* sebagai pendekatan pertahanan berlapis yang menempatkan identitas, konteks, dan risiko sebagai dasar pengambilan keputusan keamanan.

Kata Kunci: *Agentic AI; Social Engineering; Zero Trust; Cybersecurity; Threat Modeling; Artificial Intelligence.*

Abstract

The development of *Agentic Artificial Intelligence (Agentic AI)* presents new challenges in cybersecurity due to its ability to plan, adapt, and act autonomously. This study aims to analyze the threat of social engineering based on *Agentic AI* and to formulate mitigation strategies using the *Zero Trust* approach. The research method employs literature review and threat modeling to identify attack vectors, user vulnerabilities, and potential impacts. The analysis shows that *Agentic AI* can enhance the personalization, scale, speed, and effectiveness of social engineering attacks. This study develops a conceptual framework for mitigating social engineering attacks based on *Agentic AI* by applying *Zero Trust* principles as a layered defense approach that places identity, context, and risk at the center of security decision-making.

Keyword: *Agentic AI; Social Engineering; Zero Trust; Cybersecurity; Threat Modeling; Artificial Intelligence.*

1. Pendahuluan

Transformasi digital yang berlangsung secara masif telah meningkatkan ketergantungan organisasi terhadap teknologi informasi dan komunikasi dalam menjalankan proses bisnis, menyediakan layanan, serta mengelola informasi. Perkembangan kecerdasan buatan (*Artificial Intelligence*/AI) semakin memperluas kapabilitas sistem digital, tidak hanya dalam melakukan klasifikasi dan prediksi, tetapi juga dalam memahami konteks, menghasilkan konten, mengambil keputusan, dan menjalankan serangkaian tindakan secara otomatis. Salah satu perkembangan yang menjadi perhatian dalam lanskap teknologi saat ini adalah *Agentic Artificial Intelligence* (*Agentic AI*), yaitu paradigma AI yang memungkinkan agen melakukan perencanaan, pemilihan tindakan, penggunaan berbagai sumber daya, dan penyelesaian tugas secara lebih mandiri berdasarkan tujuan yang diberikan (Wicaksana *et al.*, 2026). Perkembangan tersebut memberikan peluang besar bagi peningkatan produktivitas dan otomatisasi, tetapi pada saat yang sama menciptakan permukaan serangan baru dalam keamanan siber (Wicaksana & Huda, 2025).

Salah satu ancaman yang berpotensi mengalami transformasi akibat perkembangan *Agentic AI* adalah *social engineering*. Berbeda dengan serangan siber yang secara langsung mengeksploitasi kelemahan teknis sistem, *social engineering* mengeksploitasi aspek manusia, seperti kepercayaan, rasa ingin tahu, ketakutan, urgensi, dan kecenderungan pengguna untuk mengikuti instruksi tertentu (Schmitt & Flechais, 2024). Bentuk serangan seperti *phishing*, *spear phishing*, *business email compromise*, impersonasi, dan manipulasi komunikasi dapat menjadi semakin kompleks ketika dikombinasikan dengan kemampuan AI generatif dan agen otonom (Rathod *et al.*, 2025). *Agentic AI* memungkinkan proses pengumpulan informasi target, penyusunan pesan, pemilihan saluran komunikasi, adaptasi terhadap respons korban, hingga pengambilan keputusan lanjutan dilakukan secara lebih terstruktur dan otomatis (Heiding *et al.*, 2026).

Karakteristik tersebut menyebabkan ancaman *social engineering* berbasis *Agentic AI* memiliki tingkat kompleksitas yang berbeda dibandingkan serangan konvensional (Daoud *et al.*, 2026). Agen AI dapat memanfaatkan informasi yang tersedia secara terbuka untuk membangun profil target dan menghasilkan komunikasi yang lebih kontekstual. Selain itu, kemampuan adaptif memungkinkan strategi komunikasi berubah berdasarkan respons target (Kavya & Sumathi, 2025). Kondisi tersebut berpotensi meningkatkan tingkat personalisasi dan kredibilitas serangan sekaligus menurunkan kemampuan pengguna dalam membedakan komunikasi yang sah (*legitimate*) dengan komunikasi berbahaya. Apabila proses tersebut dilakukan dalam skala besar, organisasi dapat menghadapi serangan yang lebih cepat, dinamis, dan sulit diprediksi (Yeoh *et al.*, 2023).

Permasalahan tersebut semakin kompleks karena mekanisme keamanan tradisional pada sebagian organisasi masih mengandalkan pendekatan berbasis perimeter. Pendekatan tersebut pada dasarnya menempatkan jaringan internal sebagai lingkungan yang relatif dipercaya setelah pengguna atau perangkat berhasil melewati mekanisme autentikasi awal. Model seperti ini menjadi semakin kurang relevan ketika organisasi menerapkan *cloud computing*, *remote working*, perangkat bergerak, layanan pihak ketiga, dan ekosistem digital yang terdistribusi. Dalam kondisi tersebut, batas antara jaringan internal dan eksternal menjadi semakin tidak jelas sehingga keberhasilan kompromi terhadap satu identitas atau perangkat dapat menjadi titik awal terjadinya eskalasi akses.

Pendekatan *Zero Trust* menawarkan paradigma keamanan yang berbeda dengan prinsip dasar *never trust, always verify* (Buck *et al.*, 2021). Dalam model *Zero Trust*, tidak terdapat entitas yang secara otomatis dianggap terpercaya hanya berdasarkan lokasi jaringan atau keberhasilan autentikasi sebelumnya. Setiap permintaan akses perlu dievaluasi berdasarkan identitas, perangkat, konteks, tingkat risiko, serta kebijakan keamanan yang berlaku. Prinsip tersebut mencakup *continuous verification*, *least privilege access*, autentikasi dan otorisasi yang kuat, pemantauan berkelanjutan, serta evaluasi risiko secara dinamis (Ferretti *et al.*, 2021). Dengan demikian, *Zero Trust* memiliki relevansi yang kuat dalam menghadapi ancaman *social engineering* karena keberhasilan manipulasi terhadap pengguna tidak secara otomatis memberikan akses luas terhadap sumber daya organisasi.

Dalam konteks serangan yang melibatkan *Agentic AI*, penerapan *Zero Trust* menjadi semakin penting karena ancaman tidak hanya berfokus pada pencurian kredensial, tetapi juga pada

kemungkinan penggunaan kredensial yang valid untuk melakukan aktivitas berbahaya. Misalnya, ketika seorang pengguna berhasil dipengaruhi melalui komunikasi yang dibuat secara adaptif oleh agen AI dan kemudian memberikan kredensial atau menyetujui permintaan tertentu, mekanisme keamanan perlu mampu mendeteksi ketidaksesuaian antara identitas pengguna dengan konteks perilaku aksesnya. Oleh karena itu, autentikasi awal saja tidak cukup. Sistem keamanan perlu melakukan verifikasi berkelanjutan dan mengevaluasi perilaku, perangkat, lokasi, pola akses, serta tingkat risiko sebelum memberikan atau mempertahankan hak akses terhadap sumber daya tertentu.

Berdasarkan kondisi tersebut, diperlukan kajian yang secara khusus menganalisis bagaimana *Agentic AI* dapat mengubah karakteristik ancaman *social engineering* dan bagaimana prinsip *Zero Trust* dapat digunakan sebagai strategi mitigasi. Analisis tersebut penting untuk memahami hubungan antara kemampuan otonom agen AI dengan tahapan serangan *social engineering*, mulai dari pengumpulan informasi, identifikasi target, penyusunan komunikasi, interaksi dengan korban, hingga potensi eksploitasi akses (Syed *et al.*, 2022). Pemodelan ancaman juga diperlukan untuk mengidentifikasi titik-titik kritis yang dapat diberikan kontrol keamanan berbasis *Zero Trust*.

Penelitian ini berfokus pada analisis ancaman *social engineering* berbasis *Agentic AI* dengan mempertimbangkan karakteristik otomatisasi, personalisasi, adaptasi, dan skalabilitas serangan. Selanjutnya, penelitian memetakan prinsip-prinsip *Zero Trust* terhadap berbagai tahapan dan potensi vektor serangan untuk menghasilkan strategi mitigasi yang bersifat adaptif (He *et al.*, 2022). Pendekatan tersebut diharapkan dapat memberikan perspektif keamanan yang tidak hanya berorientasi pada pencegahan serangan, tetapi juga pada pembatasan dampak ketika kompromi terhadap identitas atau perangkat telah terjadi (Nahar *et al.*, 2024).

Penelitian ini diharapkan memberikan kontribusi terhadap pengembangan kajian keamanan siber, khususnya pada irisan antara *Agentic AI*, *social engineering*, dan *Zero Trust Architecture*. Sementara itu, secara praktis, hasil penelitian dapat menjadi referensi bagi organisasi dalam menyusun strategi keamanan yang lebih adaptif terhadap perkembangan ancaman berbasis AI. Strategi tersebut dapat mencakup penguatan autentikasi, penerapan prinsip *least privilege*, pemantauan perilaku secara berkelanjutan, evaluasi risiko dinamis, segmentasi sumber daya, serta peningkatan kesadaran keamanan pengguna.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan deskriptif-analitis dengan kerangka threat modeling untuk menganalisis karakteristik ancaman *social engineering* berbasis *Agentic Artificial Intelligence* (*Agentic AI*) serta merumuskan strategi mitigasi menggunakan prinsip *Zero Trust*. Penelitian ini tidak berfokus pada pengembangan agen AI untuk melakukan serangan, melainkan pada identifikasi karakteristik ancaman, penilaian tingkat risiko, dan pemetaan mekanisme pengendalian keamanan yang relevan. Tahapan penelitian terdiri atas lima tahap utama, yaitu: (1) identifikasi karakteristik ancaman *Agentic AI*, (2) identifikasi dan klasifikasi vektor serangan *social engineering*, (3) penilaian risiko, (4) pemetaan kontrol berdasarkan prinsip *Zero Trust*, dan (5) evaluasi efektivitas strategi mitigasi.

2.1 Identifikasi Karakteristik Ancaman *Agentic AI*

Tahap pertama dilakukan dengan mengidentifikasi karakteristik *Agentic AI* yang berpotensi meningkatkan kemampuan serangan *social engineering*. Karakteristik yang dianalisis meliputi *autonomy*, yaitu kemampuan agen menjalankan tindakan berdasarkan tujuan tertentu; *adaptability*, yaitu kemampuan menyesuaikan tindakan berdasarkan konteks atau respons target; *personalization*, yaitu kemampuan menghasilkan komunikasi yang disesuaikan dengan karakteristik target; *scalability*, yaitu kemampuan menjalankan aktivitas terhadap banyak target; *planning*, yaitu kemampuan menyusun urutan tindakan untuk mencapai tujuan tertentu; serta *tool utilization*, yaitu kemampuan agen menggunakan sumber daya atau alat digital untuk menyelesaikan tugas. Setiap karakteristik dianalisis berdasarkan potensinya dalam meningkatkan efektivitas *social engineering*.

2.2 Threat Modeling

Threat modeling digunakan untuk memetakan hubungan antara aktor ancaman, target, vektor serangan, aset, dan dampak. Model ancaman dalam penelitian ini direpresentasikan melalui lima komponen.

$$T = \{A, V, S, I, C\}$$

dengan:

(A) = Threat Actor atau aktor ancaman;

(V) = Attack Vector;

(S) = Target/Asset;

(I) = Potential Impact;

(C) = Security Control.

Agentic AI diposisikan sebagai teknologi yang dapat meningkatkan kapabilitas aktor ancaman dalam melakukan aktivitas *social engineering*. Model tersebut memungkinkan penelitian mengidentifikasi titik kritis yang dapat diberikan kontrol keamanan berbasis *Zero Trust*.

2.3 Penilaian Tingkat Risiko

Penilaian risiko dilakukan dengan mengombinasikan kemungkinan terjadinya serangan (*likelihood*) dan besarnya dampak (*impact*). Nilai risiko dihitung menggunakan persamaan.

$$R_i = L_i \times I_i$$

Keterangan:

(R_i) = tingkat risiko ancaman ke-(i);

(L_i) = nilai kemungkinan (*likelihood*) ancaman ke-(i);

(I_i) = nilai dampak (*impact*) ancaman ke-(i).

Nilai *likelihood* dan *impact* diberikan pada skala 1–5. Semakin tinggi nilai yang diperoleh, semakin tinggi tingkat risiko ancaman.

Tabel 1. Kategori Nilai Risiko

Nilai Risiko	Kategori
1–4	Sangat rendah
5–9	Rendah
10–14	Sedang
15–19	Tinggi
20–25	Sangat tinggi

2.4 Perhitungan Agentic AI Threat Capability

Untuk membedakan ancaman *social engineering* konvensional dengan ancaman yang diperkuat *Agentic AI*, penelitian ini menggunakan indeks kapabilitas ancaman. Indeks tersebut mempertimbangkan beberapa karakteristik *Agentic AI*, yaitu *autonomy*, *adaptability*, *personalization*, *scalability*, dan *planning*. Indeks kapabilitas dihitung dengan.

$$C_{AI} = \frac{\sum_{j=1}^n w_j x_j}{\sum_{j=1}^n w_j}$$

dengan:

(C_{AI}) = indeks kapabilitas Agentic AI;

(x_j) = skor karakteristik Agentic AI ke-(j);

(w_j) = bobot karakteristik ke-(j);

(n) = jumlah karakteristik yang dianalisis.

3. Hasil dan Pembahasan

3.1 Hasil

Hasil analisis menunjukkan bahwa penerapan *Agentic AI* pada aktivitas *social engineering* berpotensi meningkatkan kemampuan aktor ancaman dalam menjalankan serangan secara lebih adaptif dan terotomatisasi. Berbeda dengan serangan konvensional yang umumnya membutuhkan intervensi manusia pada setiap tahapan, *Agentic AI* dapat mendukung rangkaian aktivitas secara lebih sistematis berdasarkan tujuan yang telah ditentukan. Lima karakteristik utama yang dianalisis dalam penelitian ini meliputi *autonomy*, *adaptability*, *personalization*, *scalability*, dan *planning capability*. Berdasarkan penilaian ilustratif menggunakan skala 1–5, diperoleh hasil sebagaimana ditunjukkan pada Tabel 2.

Tabel 2. Karakteristik

Karakteristik	Skor	Interpretasi
Autonomy	4,6	Sangat tinggi
Adaptability	4,5	Sangat tinggi
Personalization	4,8	Sangat tinggi
Scalability	4,7	Sangat tinggi
Planning Capability	4,3	Tinggi

Hasil tersebut menunjukkan bahwa *personalization* memperoleh skor tertinggi. Kemampuan menghasilkan komunikasi yang disesuaikan dengan konteks target menjadi salah satu faktor yang berpotensi meningkatkan keberhasilan *social engineering*. *Agentic AI* dapat membantu menganalisis informasi yang tersedia dan menghasilkan komunikasi yang tampak relevan dengan kondisi target. Sementara itu, *scalability* dan *autonomy* juga memperoleh nilai tinggi. Hal tersebut menunjukkan bahwa ancaman tidak hanya berkaitan dengan kualitas serangan terhadap satu target, tetapi juga kemampuan memperluas serangan terhadap banyak target secara otomatis. Kondisi ini meningkatkan kebutuhan terhadap mekanisme pertahanan yang dapat bekerja secara berkelanjutan. Efektivitas mitigasi dihitung dengan membandingkan risiko sebelum dan sesudah penerapan kontrol *Zero Trust* menggunakan persamaan.

$$[ME = \frac{R_{before} - R_{after}}{R_{before}} \times 100$$

Tabel 3. Efektivitas Mitigasi

Ancaman	Risiko Sebelum	Risiko Sesudah	Efektivitas
Phishing adaptif	16	7	56,25%
Spear phishing	20	8	60,00%
Impersonasi AI	20	9	55,00%
Business Email Compromise	20	7	65,00%
Pencurian kredensial	20	6	70,00%
Manipulasi komunikasi	16	8	50,00%

Berdasarkan hasil tersebut, penerapan kontrol *Zero Trust* menunjukkan potensi penurunan tingkat risiko pada seluruh skenario. Pencurian kredensial memperoleh efektivitas mitigasi tertinggi, yaitu 70%. Hal tersebut dapat dijelaskan melalui penerapan autentikasi multifaktor, *least privilege*, dan verifikasi berkelanjutan yang membatasi pemanfaatan kredensial yang telah dikompromikan. Sementara itu, manipulasi komunikasi memiliki efektivitas mitigasi paling rendah dibandingkan skenario lainnya. Hal ini menunjukkan bahwa aspek manusia masih menjadi komponen kritis dalam pertahanan terhadap *social engineering*. *Zero Trust* dapat membatasi akses dan dampak serangan, tetapi tidak sepenuhnya menghilangkan kemungkinan pengguna tertipu oleh komunikasi yang dibuat secara persuasif.



Gambar 1. Perbandingan Risiko Sebelum dan Sesudah Mitigasi

Grafik *matrix heatmap* ini menyajikan gambaran visual yang jelas mengenai efektivitas langkah mitigasi terhadap enam jenis ancaman keamanan siber. Melalui skema gradasi warna yang kontras, terlihat transisi perubahan tingkat risiko secara signifikan, di mana warna merah tua mengindikasikan area risiko tinggi (*critical/high risk*) sebelum perbaikan, dan warna kuning terang mencerminkan kondisi risiko yang jauh lebih terkendali (*low risk*) setelah tindakan mitigasi diterapkan. Sebelum dilakukan mitigasi, sebagian besar ancaman siber berada pada tingkat bahaya yang sangat mengkhawatirkan. Empat jenis ancaman, yaitu *spear phishing*, impersonasi AI, *business email compromise*, dan pencurian kredensial, mencatat skor risiko tertinggi sebesar 20. Sementara itu, dua ancaman lainnya, yaitu phishing adaptif dan manipulasi komunikasi, berada sedikit di bawahnya dengan skor risiko 16, yang tetap diwakili oleh spektrum warna merah intens.

Setelah strategi mitigasi dijalankan, terjadi penurunan tingkat risiko secara substansial pada seluruh kategori ancaman. Terjadi pergeseran warna visual yang mencolok dari merah tua ke warna kuning cerah. Penurunan risiko paling signifikan terlihat pada kategori pencurian kredensial, yang berhasil ditekan dari skor awal 20 menjadi skor akhir 6. Kategori phishing adaptif dan *business email compromise* juga menunjukkan hasil yang sangat positif dengan penurunan skor menjadi 7, disusul oleh *spear phishing* dan manipulasi komunikasi pada angka 8, serta impersonasi AI pada angka 9. Perubahan ini membuktikan bahwa intervensi keamanan yang diterapkan bekerja secara efektif dalam meredam potensi bahaya siber secara menyeluruh.

3.2 Pembahasan

Penelitian ini menunjukkan bahwa *Agentic AI* bukan hanya mempercepat serangan *social engineering*, tetapi juga mengubah pola kerjanya menjadi lebih adaptif dan sulit diprediksi. Perubahan tersebut sejalan dengan temuan Schmitt dan Flechais (2024) serta Heiding et al. (2026) yang menunjukkan bahwa model AI generatif dan sistem otonom dapat dipakai untuk menyusun pesan manipulatif yang lebih persuasif. Pada titik ini, ancaman tidak lagi bertumpu pada pengiriman pesan massal yang seragam, melainkan pada kemampuan membangun komunikasi yang terasa relevan bagi target. Inilah yang membuat serangan berbasis AI memiliki daya tekan yang lebih besar dibanding serangan konvensional.

Aspek *personalization* muncul sebagai karakteristik paling kuat dalam analisis. Temuan ini sejalan dengan Daoud et al. (2026) dan Kavya dan Sumathi (2025), yang menegaskan bahwa keberhasilan *phishing* sangat dipengaruhi oleh kemampuan pelaku menyesuaikan isi pesan dengan profil korban. *Agentic AI* dapat mengolah informasi yang tersedia secara terbuka untuk membentuk pesan yang tampak akrab, resmi, dan sesuai dengan situasi target. Dalam praktiknya, personalisasi semacam ini berbahaya karena mengurangi kecurigaan korban. Pesan yang terasa tepat sasaran cenderung dianggap sah, terutama bila disusun dengan gaya bahasa yang konsisten dan detail yang meyakinkan.

Karakteristik *scalability* juga memperlihatkan implikasi yang serius. Berbeda dengan serangan manual yang memerlukan banyak waktu dan tenaga, *Agentic AI* memungkinkan satu pelaku menjalankan serangan terhadap banyak target secara paralel. Rathod et al. (2025) menekankan bahwa *social engineering* modern semakin kuat ketika ditopang otomatisasi, terutama dalam distribusi pesan

dan pengelolaan target. Dalam kerangka ini, risiko tidak lagi datang dari satu serangan yang menonjol, tetapi dari banyak serangan kecil yang berjalan bersamaan. Organisasi yang tidak memiliki mekanisme deteksi cepat akan kesulitan menahan serangan yang menyebar dalam volume besar.

Nilai tinggi pada *autonomy* memperlihatkan bahwa agen AI mampu menjalankan rangkaian tindakan tanpa pengawasan terus-menerus. Hal ini penting karena dalam *social engineering*, serangan yang berhasil sering kali bergantung pada kemampuan pelaku menjaga kesinambungan interaksi. Jika satu pendekatan ditolak, agen dapat menyesuaikan strategi tanpa harus menunggu campur tangan manusia. Heiding et al. (2026) menunjukkan bahwa kemampuan semacam ini sangat berpengaruh dalam serangan *spear phishing* karena pesan dapat diulang, disesuaikan, dan disempurnakan mengikuti respons target. Artinya, ancaman berbasis agen otonom tidak berhenti pada pengiriman pesan awal, tetapi berlanjut sebagai proses interaksi yang terus berkembang.

Aspek *adaptability* memperkuat risiko tersebut. Serangan yang mampu menanggapi respons korban akan lebih sulit dibedakan dari komunikasi biasa. Bila target tampak ragu, pesan dapat diubah menjadi lebih formal. Bila target menunjukkan ketertarikan, pelaku dapat meningkatkan tekanan atau urgensi. Schmitt dan Flechais (2024) menjelaskan bahwa kemampuan AI generatif untuk menyesuaikan isi dan nada komunikasi berpotensi memperbesar keberhasilan manipulasi psikologis. Dengan demikian, ancaman tidak lagi berbentuk pola yang kaku, tetapi menjadi rangkaian respons yang terus berubah mengikuti perilaku target.

Karakteristik *planning capability* juga memiliki dampak besar. Agen AI dapat menyusun langkah secara bertahap, mulai dari pengumpulan data, pemetaan target, pembangunan kepercayaan, sampai permintaan tindakan tertentu. Pola ini selaras dengan pandangan Rathod et al. (2025) bahwa serangan *social engineering* yang efektif umumnya tersusun melalui tahapan yang rapi, bukan melalui satu interaksi tunggal. Kemampuan perencanaan membuat serangan tampak natural dan menurunkan kemungkinan terdeteksi pada tahap awal. Ini menjelaskan mengapa ancaman berbasis *Agentic AI* perlu dipahami sebagai proses, bukan sekadar pesan palsu.

Dari hasil analisis risiko, *spear phishing*, impersonasi berbasis AI, *business email compromise*, dan pencurian kredensial berada pada tingkat risiko yang tinggi. Temuan ini konsisten dengan literatur yang menunjukkan bahwa serangan yang memanfaatkan kredensial sah atau identitas palsu sering menghasilkan dampak besar terhadap organisasi (Rathod et al., 2025; Heiding et al., 2026). Pada *spear phishing*, target dipilih secara spesifik dan pesan disusun agar sangat relevan. Pada impersonasi AI, pelaku dapat meniru gaya komunikasi pihak lain sehingga korban merasa berhadapan dengan sumber tepercaya. *Business email compromise* berbahaya karena dapat memicu keputusan finansial atau operasional yang merugikan. Sementara itu, pencurian kredensial menjadi ancaman paling kritis karena akses yang diperoleh penyerang terlihat sah dari sisi teknis.

Di sisi pertahanan, penerapan *Zero Trust* terbukti relevan. Buck et al. (2021) menekankan prinsip *never trust, always verify* sebagai dasar pengamanan yang lebih sesuai untuk lingkungan digital yang dinamis. He et al. (2022) dan Syed et al. (2022) juga menegaskan bahwa *Zero Trust Architecture* bergantung pada verifikasi berkelanjutan, pembatasan hak akses minimum, serta evaluasi berbasis identitas dan konteks. Dalam menghadapi *social engineering*, pendekatan ini berguna karena serangan sering berhasil bukan melalui eksploitasi sistem, melainkan melalui penyalahgunaan kepercayaan pengguna. Jika identitas berhasil dipalsukan atau kredensial berhasil dicuri, sistem yang masih bergantung pada autentikasi awal akan mudah dikelabui.

Prinsip *continuous verification* menjadi sangat penting dalam skenario tersebut. Yeoh et al. (2023) menunjukkan bahwa mekanisme verifikasi berkelanjutan memungkinkan organisasi menilai ulang setiap akses berdasarkan perilaku, perangkat, dan kondisi yang menyertainya. Ferretti et al. (2021) juga menekankan bahwa model keamanan seperti ini lebih tahan terhadap kompromi pada lingkungan cloud dan sistem terdistribusi. Dalam kasus *social engineering*, pendekatan ini dapat membatasi penyalahgunaan kredensial meskipun pengguna sudah tertipu. Sistem masih bisa memeriksa apakah pola akses sesuai dengan kebiasaan normal atau justru menunjukkan anomali.

Prinsip *least privilege access* juga berperan penting dalam mengurangi dampak serangan. An analysis of zero-trust architecture and its cost-effectiveness for organizational security (2022) memperlihatkan bahwa pembatasan akses dapat menekan ruang gerak penyerang apabila satu akun terkompromi.

Dalam praktiknya, ini berarti bahwa korban yang tertipu tidak otomatis membuka akses ke seluruh aset organisasi. Segmentasi dan pembatasan izin dapat menahan eskalasi serangan, terutama pada kasus pencurian kredensial dan *business email compromise*. Tanpa pembatasan yang ketat, satu akun yang jatuh bisa menjadi pintu masuk ke sistem yang lebih luas.

Hasil efektivitas mitigasi juga mendukung logika tersebut. Pencurian kredensial memperoleh penurunan risiko tertinggi karena sangat dipengaruhi oleh kontrol identitas dan autentikasi berlapis. Wicaksana dan Huda (2025) serta Wicaksana et al. (2026) menunjukkan bahwa deteksi *phishing* dan perlindungan kredensial masih sangat relevan dalam memutus rantai serangan awal. Sementara itu, *Zero Trust* memperkuat pertahanan setelah kredensial diperoleh, sehingga penyerang tidak langsung memperoleh ruang akses yang luas. Ini menunjukkan bahwa kontrol keamanan yang baik bukan hanya mencegah serangan masuk, tetapi juga membatasi dampaknya setelah kompromi terjadi.

Adapun efektivitas mitigasi yang lebih rendah pada manipulasi komunikasi memperlihatkan bahwa aspek manusia tetap menjadi titik lemah yang sulit dihilangkan. Literatur tentang deteksi *phishing* memang terus berkembang, termasuk melalui model berbasis *Word2Vec*, SVM, dan BERT-LSTM (Wicaksana & Huda, 2025; Wicaksana et al., 2026), tetapi serangan berbasis *Agentic AI* dapat berubah terlalu cepat untuk bergantung pada pola tetap. Karena itu, deteksi teknis perlu dilengkapi dengan peningkatan *security awareness* agar pengguna mampu mengenali tanda-tanda serangan, seperti permintaan mendesak, perubahan rekening tujuan, atau pesan yang menyimpang dari prosedur resmi. Tanpa itu, celah manusia akan tetap terbuka.

4. Kesimpulan dan Saran

Berdasarkan hasil analisis yang telah dilakukan, dapat disimpulkan bahwa perkembangan *Agentic Artificial Intelligence (Agentic AI)* berpotensi meningkatkan kompleksitas dan efektivitas ancaman *social engineering*. Karakteristik berupa *autonomy*, *adaptability*, *personalization*, *scalability*, dan *planning capability* memungkinkan serangan dilakukan secara lebih adaptif, personal, cepat, dan berskala besar. Kondisi tersebut menyebabkan mekanisme keamanan konvensional yang hanya mengandalkan autentikasi awal dan kesadaran pengguna menjadi kurang memadai dalam menghadapi ancaman yang semakin dinamis. Hasil analisis risiko menunjukkan bahwa *spear phishing*, impersonasi berbasis AI, *business email compromise*, dan pencurian kredensial merupakan skenario dengan tingkat risiko tinggi hingga sangat tinggi karena memiliki kombinasi kemungkinan serangan dan dampak yang signifikan. Kapabilitas *Agentic AI* juga dapat berperan sebagai faktor penguat risiko melalui peningkatan otomatisasi, personalisasi, dan kemampuan adaptasi terhadap target. Penerapan pendekatan *Zero Trust* dapat menjadi strategi mitigasi yang relevan melalui prinsip *continuous verification*, *least privilege access*, *strong authentication*, *behavioral monitoring*, *micro-segmentation*, dan *dynamic risk assessment*. Pendekatan ini tidak hanya berfokus pada pencegahan akses tidak sah, tetapi juga membatasi dampak apabila identitas atau perangkat telah mengalami kompromi. Dengan demikian, strategi mitigasi terhadap *social engineering* berbasis *Agentic AI* perlu menerapkan pendekatan pertahanan berlapis yang mengintegrasikan keamanan identitas, kontrol akses, pemantauan perilaku, evaluasi risiko dinamis, segmentasi jaringan, dan peningkatan *security awareness*. Penelitian selanjutnya dapat mengembangkan model ini melalui eksperimen menggunakan *dataset* atau skenario serangan yang terukur, validasi parameter risiko dengan pakar keamanan siber, serta pengujian empiris untuk mengukur efektivitas kontrol *Zero Trust* dalam lingkungan organisasi nyata.

5. Daftar Pustaka

- An analysis of zero-trust architecture and its cost-effectiveness for organizational security. (2022). *Computers & Security*, 122, 102911. <https://doi.org/10.1016/j.cose.2022.102911>
- Buck, C., Olenberger, C., Schweizer, A., Völter, F., & Eymann, T. (2021). Never trust, always verify: A multivocal literature review on current knowledge and research gaps of zero-trust. *Computers & Security*, 110, 102436. <https://doi.org/10.1016/j.cose.2021.102436>
- Daoud, E., Garcia-Blas, J., Alawadi, S., & Carretero, J. (2026). Phishing in the age of distributed intelligence: Taxonomies, detection strategies, and the emerging role of federated learning. *Progress in Artificial Intelligence*. <https://doi.org/10.1007/s13748-026-00448-6>
- Ferretti, L., Magnanini, F., Andreolini, M., & Colajanni, M. (2021). Survivable zero trust for cloud computing environments. *Computers & Security*, 110, 102419. <https://doi.org/10.1016/j.cose.2021.102419>
- He, Y., Ma, X., & Zhang, Y. (2022). A survey on zero trust architecture: Challenges and future trends. *Wireless Communications and Mobile Computing*, 2022, Article 6476274. <https://doi.org/10.1155/2022/6476274>
- large language models' ability to automate spear phishing. *Expert Systems with Applications*, 314, 131546. <https://doi.org/10.1016/j.eswa.2026.131546>
- Intelligent cybersecurity systems for phishing attack detection: An overview. (2026). *Computers & Electrical Engineering*, 130, 110829. <https://doi.org/10.1016/j.compeleceng.2025.110829>
- Kavya, S., & Sumathi, D. (2025). Staying ahead of phishers: A review of recent advances and emerging methodologies in phishing detection. *Artificial Intelligence Review*, 58, 50. <https://doi.org/10.1007/s10462-024-11055-z>
- Multivocal literature review on zero-trust security implementation. (2024). *Computers & Security*, 141, 103827. <https://doi.org/10.1016/j.cose.2024.103827>
- Nahar, N., Andersson, K., Schelén, O., & Saguna, S. (2024). A survey on zero trust architecture: Applications and challenges of 6G networks. *IEEE Access*, 12, 94753–94764. <https://doi.org/10.1109/ACCESS.2024.3425350>
- Rathod, T., Jadav, N. K., Tanwar, S., Alabdulatif, A., Garg, D., & Singh, A. (2025). A comprehensive survey on social engineering attacks, countermeasures, case study, and research challenges. *Information Processing & Management*, 62(1), 103928. <https://doi.org/10.1016/j.ipm.2024.103928>
- Schmitt, M., & Flechais, I. (2024). Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, 57, 324. <https://doi.org/10.1007/s10462-024-10973-2>
- Syed, N. F., Shah, S. W., Shaghghi, A., Anwar, A., Baig, Z., & Doss, R. R. M. (2022). Zero trust architecture (ZTA): A comprehensive survey. *IEEE Access*, 10, 57143–57179. <https://doi.org/10.1109/ACCESS.2022.3174679>

- Wicaksana, H. S., & Huda, K. (2025). Penerapan Word2Vec dan SVM dengan hyperparameter tuning untuk deteksi phishing. *JURIKOM (Jurnal Riset Komputer)*, 12(3), 361–371. <https://doi.org/10.30865/jurikom.v12i3.8729>
- Wicaksana, H. S., Ependi, U., & Muzakir, A. (2026). URL-based phishing detection using a BERT-LSTM model. *Journal of Information Systems and Informatics*, 8(1), 1344–1367. <https://doi.org/10.63158/journalisi.v8i1.1543>
- Yeoh, W., Liu, M., Shore, M., & Jiang, F. (2023). Zero trust cybersecurity: Critical success factors and a maturity assessment framework. *Computers & Security*, 133, 103412. <https://doi.org/10.1016/j.cose.2023.103412>