Artificial intelligence is changing social engineering from a one-time impersonation of authority into a sustained simulation of relationship. Children, seniors, employees, and isolated adults may face different-looking threats, but the attack pattern is increasingly the same: adversaries exploit trust, belonging, and connection to suppress verification, create leverage, and establish persistent control. *Belonging Is the Attack Surface* maps this human-layer threat using familiar cybersecurity concepts and proposes layered defenses for families, organizations, and communities. Its portable resilience model is: **Critical thinking + Guardrails + Safe disclosure = Resilience.**

Belonging Is the Attack Surface

AI-enabled social engineering against the human trust boundary

Artificial intelligence has changed the quality of social engineering. *Gone are the days of easily identifiable, awkwardly worded emails.* Now AI can generate fluent phishing messages, adapt tone to a target, impersonate voices, create synthetic identities, and maintain a convincing conversation at a scale that previously required large amounts of human labor. But a narrow focus on phishing, deepfakes, and malware misses a more consequential shift.

The underlying driver is not exotic. People engage online for the same reasons they always have: boredom seeking relief, and the basic human need to be noticed. Those are not vulnerabilities to be patched — they are baseline human behavior. That behavior runs on a specific mechanism: unpredictable social reward — a reply, a notification, an unexpected connection — drives dopamine-mediated reward-seeking more strongly than reward delivered on a predictable schedule, the same variable-ratio principle that makes slot machines and many consumer platforms effective by design. [11] What has changed is that AI has collapsed the cost of counterfeiting human-like behavior: an adversary no longer needs charisma, patience, or luck to manufacture the feeling of being understood at scale. That is the actual attack surface this paper is about.

The target is no longer only a credential, payment, endpoint, or privileged account. It is increasingly a person’s trust, belonging, identity, and willingness to keep a dangerous interaction private—especially when a vulnerable individual, whether a child, senior, isolated adult, or person in crisis, is reachable through a platform.

This is not a claim that every harmful online interaction is caused by isolation, boredom, or a desire for acceptance. Nor is it an attempt to turn ordinary friendship or online community into a threat indicator. People seek connection because connection is normal and necessary. The security problem arises when an adversary learns to imitate connection well enough to redraw a target's trust boundaries: elevate the attacker, demote protective adults or institutions, normalize secrecy, and make verification feel like betrayal.

That is the operating logic behind a widening class of attacks that includes AI-enabled fraud, romance and investment scams, impersonation, sextortion, coercive online communities, and the extreme cases represented by 764-type violent networks. The mechanisms differ, but the attack surface is related: the normal human need to be noticed, accepted, and included.

The shift: from authority impersonation to relationship simulation

Traditional social engineering often asks a target to accept a claim:

- "I am your bank."
- "I am IT support."
- "I am your supervisor."
- "Your account is locked."
- "Pay this invoice immediately."

The attacker borrows institutional authority, creates urgency, and hopes the target acts before verifying.

Relationship simulation asks the target to accept something deeper:

- "I understand you."
- "I am like you."
- "You can tell me what is really happening."
- "This is private."
- "You owe me."
- "We are the people who understand you."

The first kind of attack seeks a transaction. The second seeks a relationship—or, more accurately, a credible simulation of one. Its objective is not merely a click. It is durable influence.

CISA describes social engineering as using human interaction to obtain or compromise information about an organization or its systems. That definition remains correct. What deserves greater attention is how the interaction itself has become a scalable attack channel: attackers can use AI-generated text, false personas, and adaptive scripts to build trust over time rather than merely exploit a momentary lapse. [1][2]

This does not replace traditional phishing. It extends it. A credential-harvesting email may be the initial access vector; a direct message from a convincing peer, executive, recruiter, investor, romantic interest, or trusted community member may be the persistence mechanism.

The relational attack lifecycle

Security professionals already recognize that successful intrusions proceed through stages. The same logic applies when the target is a person’s relational decision-making system.

Security phase	**Conventional enterprise attack**	**Relational exploitation attack**
Reconnaissance	Employee roles, vendors, org chart, breached data	Interests, social graph, posts, routines, language, communities, visible distress
Target selection	Privileged user, finance role, high-value account	Reachable person whose trust, access, identity, money, or social position can be exploited
Initial access	Phishing, spoofed call, malicious attachment	DM, game chat, group invitation, romance approach, compromised friend account
Trust establishment	Logo, familiar executive language, authority claim	Flattery, common interest, apparent peer status, shared vulnerability, persistent attention
Authentication bypass	Credential theft, MFA fatigue, help-desk manipulation	“Prove you trust me,” secrecy, loyalty tests, staged intimacy, image or information capture
Privilege escalation	Higher permissions, access tokens, administrator role	Private or encrypted channels, reduced contact with trusted adults, escalating compliance

Collection	Credentials, PII, source code, payment data	Images, recordings, location, passwords, disclosures, compromising material
Command and control	Persistent remote access, covert channels	Repeated contact, cross-platform migration, blackmail, group pressure, threat cycles
Impact	Fraud, extortion, ransomware, data theft	Financial loss, exploitation, reputational harm, coerced self-harm, lasting trauma
Incident response	Containment, evidence preservation, reporting, recovery	Stop further compromise, preserve evidence, report, restore trusted support, protect the victim

The analogy has limits. Human coercion is not a technical breach, and the victim is not a compromised machine. The purpose of the model is operational clarity: it shows why a one-time safety warning is inadequate once an attacker has established rapport, isolated the target, acquired compromising material, or created a continuing threat channel.

A particularly important technical translation is this: **coercion is persistence.** In an enterprise intrusion, persistence enables continued access after initial compromise. In relational exploitation, compromise—an image, secret, account, recording, payment, or apparent rule violation—can enable continued control. The attacker’s goal is not simply to acquire material; it is to create a durable mechanism for repeated compliance.

This is not merely rhetorical. A trauma bond — an emotional attachment to the person exercising control, reinforced through cycles of threat, tension, and intermittent reconciliation — can form during the threat phase and helps explain sustained compliance long after continued contact stops looking rational from the outside. [9] A related fawn response — appeasement directed at the person holding the threat, as a survival strategy — is well documented in trauma literature and offers a plausible mechanism for escalating cooperation rather than disengagement. [10] Neither construct is drawn from a study of this exact population; both function here as structural analogies for a coercion dynamic that runs on the same intermittent-reinforcement logic security teams already recognize from other manipulation contexts. That intermittent-reinforcement logic is not only structurally similar — it is the same dopamine-mediated mechanism described above, and it plausibly explains why the relief of reconciliation with the person exercising control becomes something the nervous system comes to seek, not merely tolerate. [11]

Why AI changes the threat model

AI does not invent deception, grooming, fraud, or coercion. It changes the economics and operational quality of each stage.

Reconnaissance compression

People leave extensive public and semi-public behavioral traces: posts, comments, follows, usernames, videos, interests, affiliations, language patterns, locations, professional roles, and visible relationship networks. An attacker has long been able to use this material manually. AI can lower the cost of collecting, summarizing, sorting, and converting it into an approach strategy.

The critical risk is not a magical system that infallibly “knows” a target. It is that adversaries can cheaply generate a plausible hypothesis about what invitation, tone, concern, authority claim, shared interest, or urgency cue is most likely to receive a response.

Relationship simulation at scale

A capable manipulator once needed writing skill, patience, memory, and time to sustain many tailored conversations. AI can now assist with language variation, translation, age- or community-coded phrasing, tone matching, follow-up questions, apparent empathy, and response suggestions.

This lowers the threshold for operating a false persona. It also increases the number of parallel conversations an offender can sustain without making the usual errors of repetition, poor grammar, or inconsistent tone.

Government warnings already recognize the narrower version of this point: the FBI has stated that criminals use AI-generated text to make messages more believable in social engineering, spear phishing, and financial fraud. [2] The broader working synthesis is that believable text can be used not only to make a target click, but to make a target remain.

Synthetic identity and social proof

The practical risk is no longer limited to a badly edited profile picture or an implausible email domain. AI can help create images, voices, video fragments, profiles, apparent chat histories, endorsements, and other artifacts that together create synthetic social proof.

No single artifact must be perfect. A convincing interaction can emerge from accumulation:

The relevant security question is therefore not simply, “Is this image authentic?” It is: *What combination of identity claims, cross-platform cues, apparent shared contacts, and emotionally resonant interaction is enough to suppress verification?*

A distinct and increasingly significant case does not require a relationship, or even prior contact, at all. AI image generation can convert an ordinary, publicly available photo of a real, identifiable minor into fabricated sexually explicit content, which is then used directly for extortion. The FBI's Internet Crime Complaint Center has warned specifically of this pattern since 2023. [3] This case sits outside the relational attack lifecycle described above: there is no reachability-through-relationship, no recognition, no secrecy built through conversation. Reconnaissance and compromise-material generation collapse into a single automated step, and the lifecycle can begin at threat. Any threat model built on the relational lifecycle should treat this as a separate entry point, not a variant of it.

Adaptive escalation

A conventional phishing template is relatively static. AI enables a more iterative loop:

1. Generate an opening message.
2. Observe the response, hesitation, stated interest, or vulnerability cue.
3. Tailor a follow-up.
4. Test a small boundary.
5. Escalate, retreat, flatter, or threaten according to resistance.
6. Move the conversation to a less visible channel.
7. Reuse successful patterns at scale.

This is not an assertion that AI independently forms malicious intent. It is an account of how human adversaries can use low-cost adaptive tools to improve targeting and sustain interaction.

FBI warnings regarding 764 and similar violent online networks describe manipulation and coercion of vulnerable and underage people to produce child sexual abuse material and other harmful content. [4] Those networks represent an extreme endpoint, not a model for all online harm. But they show how progressive testing, isolation, compromise, and threats can form a continuing control system.

Belonging as a trust-boundary problem

Security professionals understand trust boundaries: places where a system grants access, accepts an identity claim, or permits one component to affect another.

Belonging can be understood in parallel. A person carries an internal trust map: who is safe, who has authority, who may ask for help, whose approval matters, what must remain private, and which relationships can be questioned.

A relational attacker tries to redraw that map.

Attacker objective	**Trust-boundary manipulation**
Elevate the attacker	“I understand you better than anyone else.”
Lower trust in protectors	“Your parents, friends, boss, or school would not understand.”
Normalize secrecy	“This is only between us.”
Make verification disloyal	“If you check with someone else, you are betraying me.”
Create dependency	“You need this group, relationship, opportunity, or approval.”
Create leverage	“If you stop, I will expose what you shared.”

The trust-boundary problem also has a version with no adversary at all. AI companion products — chatbots designed for conversation, role-play, and emotional support — are used by a majority of teens, and are commercially optimized for the same need this paper treats as an attack surface: to be noticed, validated, and never contradicted. [5] No one needs to attack a trust boundary that a product is already paid to move on its own. This is not a claim that AI companionship is inherently harmful; it is a claim that the underlying mechanism — belonging simulated at scale, monetized by engagement — is identical to the one this paper treats as adversarial, minus the adversary. Security teams

evaluating human-layer risk should not assume an attacker is required for the mechanism to cause harm.

This helps explain why “user awareness” is too weak a term. An employee who receives a suspicious invoice can often consult a colleague or verify through a known channel. A child, isolated adult, or employee entangled in a simulated relationship may experience verification as dangerous, humiliating, or disloyal.

The attack is therefore not only against judgment. It is against critical thinking, the guardrails surrounding the relationship, and the target’s ability to access safe disclosure and a second opinion.

Defense in depth for the human layer

The standard security answer to a complex threat is not a single control. It is defense in depth. The same principle applies here.

The human-readable form of this defense model is:

Critical thinking + Guardrails + Safe disclosure = Resilience

This formula does not replace technical controls or established security disciplines. It clarifies how the controls work together at the human trust boundary.

Core countermeasure	**Security function**	**Human-layer implementation**
Critical thinking	Resist deception before action	Out-of-band confirmation, known-channel callbacks, source checks, “What is missing?” questions, recognition of urgency, secrecy, isolation, and boundary testing
Guardrails	Reduce exposure, opportunity, and isolation; enable early detection	Privacy defaults, restricted DMs, age-appropriate accounts, identity and payment controls, shared-space and device norms, monitoring, account hygiene, trusted-adult access, reporting pathways
Safe disclosure	Enable containment, recovery, and learning after concern or compromise	Blame-minimizing escalation, evidence preservation, reporting, victim-centered support, account recovery, counseling, manager or family involvement, post-incident review without punishment

No element is sufficient alone. Critical thinking can fail under pressure. Guardrails can be bypassed. A person may recognize danger yet remain silent because of shame, fear, or a credible threat. Together, these layers make manipulation harder to initiate, harder to sustain, and easier to interrupt.

The same mapping can be expressed in defense-in-depth terms:

Defensive layer	**Security function**	**Human-layer analog**	**Core countermeasure**
Prevent	Reduce exposure and attack opportunity	Privacy defaults, restricted DMs, age-appropriate accounts, shared-space norms, account hygiene	Guardrails
Detect	Identify suspicious activity early	Notice secrecy, escalating pressure, isolation, sudden fear, abrupt behavioral change, unusual private contact	Guardrails
Verify	Resist deception before action	Out-of-band confirmation, known-channel callbacks, “What is missing?” questions, source checks	Critical thinking
Contain	Limit attacker persistence and spread	Stop sending material or money; preserve evidence; report; block once evidence and reporting steps are underway	Safe disclosure
Recover	Restore safe operation and support	Shame-free disclosure, counseling, account recovery, family support, victim-centered reporting	Safe disclosure
Learn	Improve future resilience	Scenario practice, relational threat literacy, updated norms, post-incident review without blame	Critical thinking + Guardrails + Safe disclosure

The human-layer model should not become an excuse to shift responsibility from platforms, companies, or governments to individual families. It is a complement to systemic controls, not a substitute for them. Platforms have obligations to improve privacy defaults, detect abuse, disrupt fraudulent accounts, preserve reporting pathways, and reduce recommender-system amplification of harmful contact. Organizations must strengthen identity verification, controls around payments and sensitive requests, and incident response. Law enforcement and policymakers retain separate responsibilities for investigation, deterrence, and victim support.

Still, when a message reaches a target, an internal and relational defense is necessary. Technical controls can reduce exposure. They cannot fully replace a person’s ability to pause, verify, retain contact with trusted others, and disclose a problem without shame.

Incident response: protect the person, preserve the evidence

The response sequence for coercive contact is similar in spirit to incident response but must be victim-centered.

1. **Stop further compromise.** Do not send more images, money, credentials, personal information, or “proof.” Do not attempt to appease or negotiate with the coercive actor.
2. **Preserve evidence.** Save usernames, account URLs, messages, timestamps, threats, payment requests, and relevant screenshots. Do not delete the profile or conversation before needed information is preserved.
3. **Report and contain.** Report through the platform and the appropriate authorities. Isolate the offending contact—block, mute, or leave the specific channel—once evidence has been preserved and reporting has begun, rather than disconnecting the device itself.
4. **Restore support.** Bring in a trusted adult, supervisor, counselor, legal adviser, victim-support organization, or other appropriate support. The immediate goal is to end isolation—not to assign blame.
5. **Restore safe disclosure and learn without blame.** Determine what controls failed, what support was missing, and how verification, guardrails, reporting, and support can improve. Avoid treating the victim’s understandable response to coercion as the root cause.

NCMEC and FBI sextortion guidance emphasizes that victims should not comply with a blackmailer, should preserve relevant account and message information, and should report suspected exploitation. [6][7]

For organizations, this model supports a needed refinement in security culture. An employee who reports a coercive or embarrassing interaction should encounter a safe escalation path, not automatic humiliation. A workforce trained only to fear punishment will be easier to silence after an error or compromise.

One pattern this model does not cleanly cover deserves a direct note: deepfake image generation used by minors against their own peers, for humiliation rather than extortion. NCMEC has documented a sharp rise in these cases. [8] The lifecycle model above assumes an adversary distinct from the victim's peer group; peer-perpetrated deepfake abuse breaks that assumption, and organizations building on this framework — schools especially — should treat it as a related but structurally distinct risk rather than force-fitting it into the same response playbook.

Implications for security leaders

Expand threat modeling

Threat models should include relationship simulation, cross-platform trust building, AI-assisted impersonation, and persistence through coercive leverage—not only malicious links and credential theft.

Treat verification as a social practice

Out-of-band verification is not merely a technical procedure. It must be culturally safe. People need permission to slow down, call a known number, ask a colleague, involve a manager, or disclose an embarrassing interaction without being treated as incompetent.

Treat safe disclosure as a security control

Safe disclosure is not only a wellbeing measure or an after-incident courtesy. It is an early-warning and containment control. When people can report coercive, suspicious, or embarrassing contact without automatic humiliation or punishment, organizations learn about attacks sooner, preserve better evidence, and limit persistence.

Conversely, punitive culture strengthens an attacker’s leverage. If an employee expects blame for clicking, replying, sharing information, or becoming emotionally entangled, the attacker’s threat of exposure has more power. Organizations should therefore define confidential, blame-minimizing escalation paths; train managers to respond calmly; and separate immediate safety and containment from later accountability review.

Train for patterns, not static indicators

The next generation of social engineering may have correct grammar, accurate personal details, and realistic voices. Training must move beyond spotting errors and toward recognizing patterns.

For general audiences, describe these plainly as **relationship red flags**. In security programs, **relational threat literacy** remains a useful term for the capability to identify, interpret, and respond to these patterns.

Pattern	What it looks like
Urgency	Pressure to act, respond, or decide immediately, with no room to verify
Secrecy	An expectation that the interaction stay hidden from trusted others
Isolation	Discouragement—explicit or implied—from checking with family, colleagues, or supervisors
Unusual requests	Money, credentials, images, or access not normally asked for through this channel
Boundary testing	Small asks that escalate step by step, each slightly further than the last
Payment pressure	Requests for money, gift cards, cryptocurrency, or financial access, especially urgent or unusual ones
Attempts to prevent consultation	Discouraging a second opinion, out-of-band verification, or involving another trusted person

Include families and communities

For minors and vulnerable populations, security cannot end at the enterprise perimeter. Parents, educators, youth-serving organizations, and community partners need accessible models of how manipulation evolves from contact to control. This is not peripheral social work. It is prevention at the human layer.

Use AI for defense, carefully

AI can support safe scenario-based learning: generating fictional messages for employees, parents, or youth to analyze for urgency, secrecy, impersonation, and missing verification steps. This is a proposed practice, not a proven prevention program.

It should supplement—not replace—human supervision, platform safeguards, security controls, and professional intervention when a person is at risk.

Conclusion

AI will make deception more fluent, personal, and persistent. It will improve the adversary’s ability to imitate authority, familiarity, and increasingly, relationship.

The appropriate response is not fear of all online connection or a belief that people can be protected through filtering alone. It is a broader security model: protect technical systems, but also protect the human trust relationships through which systems, identities, money, and safety are increasingly compromised.

Belonging is not a weakness. It is a normal human need and a source of resilience when it is real, accountable, and connected to protective communities. But when an attacker can simulate belonging, isolate a target from independent verification, and convert shame into silence, belonging becomes an attack surface.

The security task is to make that surface harder to exploit:

Critical thinking + Guardrails + Safe disclosure = Resilience

Critical thinking creates a pause. Guardrails reduce exposure and isolation. Safe disclosure ensures that a target can access help before a harmful interaction becomes a durable channel of control.

References

[1] Cybersecurity and Infrastructure Security Agency. (2021, February 1). *Avoiding social engineering and phishing attacks.* https://www.cisa.gov/news-events/news/avoiding-social-engineering-and-phishing-attacks

[2] Federal Bureau of Investigation. (2024, May 9). *FBI warns of increasing threat of cyber criminals utilizing artificial intelligence.* FBI San Francisco. https://www.fbi.gov/contact-us/field-offices/sanfrancisco/news/fbi-warns-of-increasing-threat-of-cyber-criminals-utilizing-artificial-intelligence

[3] Federal Bureau of Investigation, Internet Crime Complaint Center. (2023, June 5). Malicious actors manipulating photos and videos to create explicit content and sextortion schemes. https://www.ic3.gov/PSA/2023/psa230605

[4] Federal Bureau of Investigation. (2025, March 6). *Violent online networks target vulnerable and underage populations across the United States and around the globe.* https://www.fbi.gov/investigate/cyber/alerts/2025/violent-online-networks-target-vulnerable-and-underage-populations-across-the-united-states-and-around-the-globe

[5] Common Sense Media. (2025, July). Nearly 3 in 4 teens have used AI companions, new national survey finds. https://www.commonsensemedia.org/press-releases/nearly-3-in-4-teens-have-used-ai-companions-new-national-survey-finds

[6] National Center for Missing & Exploited Children. (n.d.). *Sextortion.* https://www.ncmec.org/theissues/sextortion

[7] Federal Bureau of Investigation. (2024, January 12). *The financially motivated sextortion threat.* https://www.fbi.gov/news/stories/the-financially-motivated-sextortion-threat

[8] National Center for Missing & Exploited Children. (2025). The deepfake dilemma: New challenges protecting students, confidentiality. https://www.missingkids.org/blog/2025/the-deepfake-dilemma-new-challenges-protecting-students-confidentiality

[9] Speck, P. M., & Patrician, P. A. (2019). A concept analysis of trauma coercive bonding in the commercial sexual exploitation of children. Journal of Pediatric Nursing, 46, 48–54. https://doi.org/10.1016/j.pedn.2019.02.030

[10] Walker, P. (2003, January/February). Codependency, trauma and the fawn response. The East Bay Therapist. https://www.pete-walker.com/codependencyFawnResponse.htm

[11] Özüçetin, B. (2026). Social approval, reward learning, and the neurobiological basis of social media use: A narrative review. Journal of Consultation Liaison Psychiatry, 2(1), 10–21. https://doi.org/10.71350/jclp.27

Source note

The relational attack lifecycle, the “coercion is persistence” framing, the trust-boundary mapping, and the model **Critical thinking + Guardrails + Safe disclosure =**

Resilience are analytical frameworks developed in this paper. The cited materials provide factual context and response guidance; they do not independently validate every element of the framework.