Trust and Verification in Digital Impersonation and Cyber Deception: A Human Factors Review and Research Agenda

Calvin Nobles[1] and Samson Quaye[2]

[1]School of Cybersecurity and Information Technology, University of Maryland Global Campus, Adelphi, MD, USA

[2]Center for Cybersecurity and Forensic Education (C2SAFE), Illinois Institute of Technology, Chicago, Illinois, USA

drcalvinnobles@gmail.com
squaye@hawk.illinoistech.edu

Abstract: Digital impersonation exploits trust, verification behavior, and decision-making under uncertainty. This structured narrative review synthesizes selected literature on cybersecurity, social engineering, trust in automation, AI-mediated communication, and synthetic media. Four themes organize the analysis: trust calibration, cognitive biases and social influence, verification failure, and human-centered defenses. The literature suggests that susceptibility reflects interactions among credibility cues, authority pressure, urgency, cognitive demands, and organizational practices. Generative AI can reproduce familiar communication cues, challenging verification based on appearance, fluency, or sensory familiarity. However, evidence from conventional phishing and automation research does not establish the magnitude of these effects in every AI-enabled setting. The review proposes research priorities addressing trust measurement, verification fatigue, organizational culture, and interactive deception. Its central implication is that effective defense requires usable verification procedures and organizational support alongside technical detection and training.

Keywords: Digital impersonation, Cyber deception, Trust calibration, Verification behavior, Human factors, AI-enabled deception

1. Introduction

Artificial intelligence (AI) expands the ways attackers can impersonate trusted individuals and organizations. Deepfake video, synthetic speech, and generated text can reproduce cues associated with legitimate communication. These capabilities extend established phishing, pretexting, and business email compromise practices, making digital impersonation a problem of human judgment as well as technical security.

Impersonation exploits familiar names, voices, faces, institutional branding, and communication patterns. Synthetic political video can generate uncertainty and undermine trust in information (Vaccari & Chadwick, 2020). Research on audio deepfakes also identifies limitations in human detection, although performance depends on the samples and conditions evaluated (Warren et al., 2024). These findings caution against treating perceptual realism as reliable evidence of authenticity.

Technical controls remain essential, but their effectiveness depends partly on how users interpret warnings and implement verification procedures. Employees may overlook risk when workload, urgency, or organizational hierarchy discourages scrutiny. Consequently, security outcomes reflect how system design, cognitive demands, and workplace practices interact.

Trust calibration offers a useful analytical foundation. Lee and See (2004) describe appropriate reliance on automation as a relationship between trust and system capability. Applied to impersonation, this perspective highlights whether confidence is justified by available evidence. Excessive confidence can encourage unsafe compliance, whereas indiscriminate distrust can impede legitimate communication.

Social engineering exploits authority, urgency, familiarity, and interpersonal obligation (Heartfield & Loukas, 2015; Siddiqi et al., 2022). AI may amplify these mechanisms by producing convincing content at scale, but empirical investigation is needed to understand the effects of specific technologies and contexts.

This review examines how human factors shape trust, verification, and susceptibility to digital impersonation. It synthesizes four themes and develops a research agenda for verification design, synthetic impersonation, and organizational culture. The central argument is that defenses should make reliable verification feasible within ordinary work, rather than depend on sustained individual vigilance.

2. Review Approach

This paper uses a structured narrative review to examine how human factors influence trust, verification behavior, and susceptibility to digital impersonation. Sources were selected for conceptual or empirical

relevance to phishing, social engineering, synthetic media, AI-mediated communication, trust, persuasion, and verification. Foundational scholarship provides theoretical grounding; newer studies inform discussion of emerging deception capabilities. Policy and practitioner publications provide governance context rather than evidence of intervention effectiveness.

The review excludes work confined to detection algorithms, cryptographic authentication, or technical controls without a behavioral or organizational component. It does not claim exhaustive searching, formal risk-of-bias appraisal, or compliance with the PRISMA reporting framework (Page et al., 2021). The authors do not report a reproducible search and screening record, limiting assessment of completeness and independent replication. Conclusions should therefore be interpreted as an integrative conceptual synthesis, not pooled estimates or a comprehensive assessment of the field.

The synthesis is organized around trust calibration, cognitive and social influence, verification failure, and human-centered defenses. Proposed relationships and research priorities are distinguished from established empirical findings; generalizations from conventional phishing to emerging AI-enabled contexts remain provisional.

3. Conceptual Background

Digital impersonation involves a false, stolen, or manipulated identity used to deceive a target. It can enable fraudulent payments, information disclosure, access to malicious links, or unauthorized approvals. Examples include spoofed accounts, executive impersonation, synthetic voices, deepfake video, and fabricated identities.

Here, cyber deception refers to malicious efforts to mislead people through digital communication or systems. Impersonation is a subset that exploits beliefs about identity. Sender names, logos, tone, and familiar requests may support routine communication, but they do not, by themselves, establish authenticity.

A human factors perspective examines how perception, attention, workload, social expectations, and system design shape security decisions. Social engineering research links deception to both persuasive mechanisms and organizational processes (Heartfield & Loukas, 2015; Krombholz et al., 2015). This framing locates vulnerability within the broader work system rather than solely within individual users.

Trust denotes willingness to rely on a person or system under uncertainty. Verification comprises actions used to establish authenticity or authorization, such as contacting a known number, using authenticated channels, or escalating an unusual request. Verification quality depends on the independence and reliability of the evidence obtained.

The objective is proportionate trust supported by verification appropriate to the consequences of an action (Lee & See, 2004; Vishwanath et al., 2011). Table 1 summarizes the core concepts. Figure 1 presents the proposed relationships among impersonation cues, human factors, trust, and verification. It is a conceptual synthesis, not an empirically validated causal model.

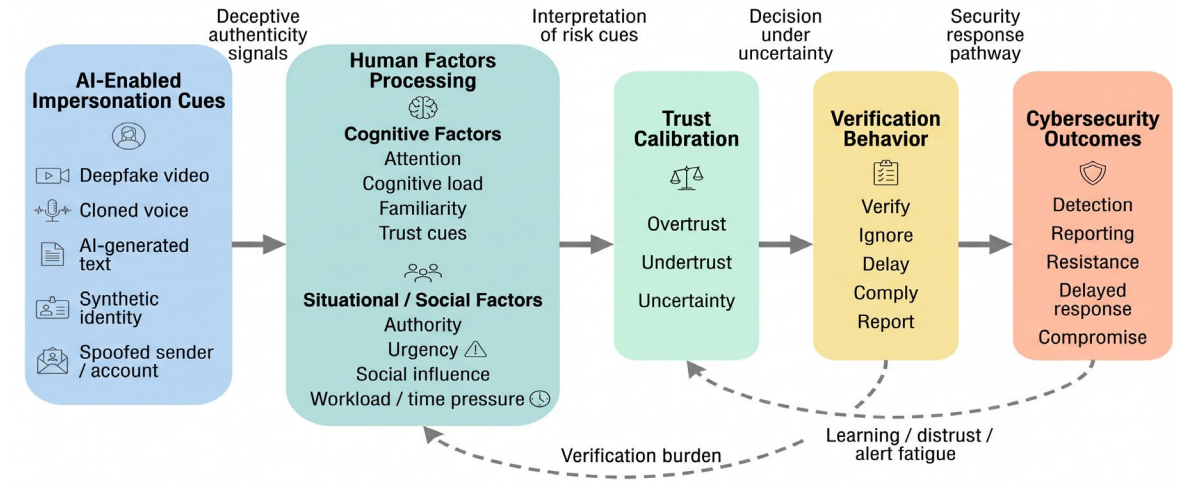

Figure 1: Conceptual model of trust and verification in AI-enabled digital impersonation

Table 1: Core Concepts in Digital Impersonation and Human Factors

Concept	Operational Definition	Human Factors Relevance
Digital impersonation	Use of a false, stolen or manipulated identity to deceive a target	Exploits trust in names, voices, faces, accounts, roles and institutions
Cyber deception	Deliberate use of digital communication or systems to mislead users	Depends on how users perceive risk, interpret cues and make decisions
Trust calibration	Alignment between user trust and actual reliability or risk	Overtrust may increase unsafe compliance; undertrust may impede legitimate work
Verification behavior	Actions taken to confirm whether a message, identity or request is authentic	Can fail under time pressure, cognitive load, uncertainty or authority pressure
Social engineering	Psychological manipulation used to influence user behavior	Uses persuasion, authority, urgency, familiarity and social pressure
Synthetic media	AI-generated or manipulated audio, video, image or text	Makes familiar trust cues easier to imitate and harder to verify

4. Trust Calibration and Overtrust

4.1 Trust Calibration in Cybersecurity

Trust calibration concerns the correspondence between confidence and the reliability of the evidence supporting reliance. In impersonation settings, users may infer legitimacy from a familiar sender, professional language, or a recognizable voice. These cues become unreliable when attackers reproduce them without possessing the claimed identity or authority.

AI-generated communication can separate apparent credibility from genuine provenance. Fluent text and realistic audiovisual content may resemble legitimate interaction while providing little independent evidence of identity. The relevant design question is therefore how users can access trustworthy evidence without undertaking an impractical investigation of every message.

4.2 Overtrust, Undertrust and Verification Fatigue

Overtrust can encourage fraudulent approvals or information disclosure. Undertrust may obstruct legitimate work and reduce willingness to use helpful security tools. Verification fatigue is considered here a proposed consequence of repeated checking demands; its prevalence and relationship to performance require direct measurement rather than assumption.

Trust in automation research suggests that reliance depends on users' understanding of system capabilities and limitations (Lee & See, 2004). In cybersecurity, an unflagged message should not be interpreted as proof of safety. Similarly, frequent, poorly explained, or hard-to-act-on warnings may undermine reliance. These applications of automation theory require testing in specific deception contexts.

4.3 Synthetic Media and Trust Uncertainty

Synthetic media can make fabricated content plausible while also weakening confidence in authentic evidence (Chesney & Citron, 2019; Vaccari & Chadwick, 2020). Detection difficulty should not be treated as uniform across people, modalities, or tasks. A defensible approach combines technical evidence with independent verification and explicitly communicates uncertainty.

Trust calibration is consequently a design and governance concern. Systems should distinguish identity evidence from persuasive presentation, and procedures should make additional checks proportionate to the potential harm of the requested action.

5. Cognitive Biases and Social Influence

5.1 Cognitive Shortcuts in Deception

People use cognitive shortcuts to manage high volumes of communication. Sender identity, tone, formatting, and contextual familiarity support efficient decisions but can be imitated. These strategies are not inherently irrational; their reliability depends on the environment in which they are used.

Vishwanath et al. (2011) link phishing susceptibility to information processing and how people interpret message cues. This finding supports attention to how users evaluate a request, rather than merely whether they recognize a suspicious feature. In AI-enabled contexts, grammatical quality alone provides insufficient grounds for judging legitimacy.

5.2 Authority, Urgency and Compliance

Attackers may invoke executives, institutions, or technical support personnel to establish apparent authority. Urgent deadlines can further discourage deliberation. Together, these factors can make verification seem inconsistent with responsiveness or compliance, especially when employees lack clear permission to challenge unusual requests.

Cialdini and Goldstein (2004) describe social influence processes underlying compliance and conformity. Social engineering research applies related concerns to psychological manipulation and routine trust relationships (Krombholz et al., 2015; Siddiqi et al., 2022). However, these foundations do not establish that every influence mechanism has equivalent effects across AI-generated messages, voice calls, and interactive agents.

5.3 Familiarity and Truth Bias

Impersonation can borrow trust from an existing relationship. A recognizable name, expected invoice, or familiar communication style may lower suspicion without confirming the source. The persuasive value of familiarity should therefore be distinguished from the evidentiary value of authentication.

Interpersonal deception theory situates deception within interactive communication and relationships (Buller & Burgoon, 1996). Expectations of honest communication can shape how people interpret ambiguous messages. Extending this account to AI-generated interaction is theoretically useful, but dedicated studies are needed to examine the effects of fluency, confidence, and conversational adaptation.

5.4 Implications for Human Factors Design

Design responses should reduce dependence on weak cues and support deliberate checks for consequential actions. Clear warnings, accessible verification channels, and explicit permission to question authority can make secure behavior compatible with ordinary work.

6. Verification Failure

6.1 Why Verification Breaks Down

Verification fails when users omit checks, accept inadequate evidence, or consult a channel controlled by the attacker. Such failure may reflect workload, unclear procedures, or social pressure rather than a lack of security knowledge. Phishing and pretexting research identifies urgency and habitual message processing as relevant vulnerabilities (Vishwanath et al., 2011; Workman, 2008).

A request that resembles routine work can direct attention toward task completion rather than source validation. This distinction matters because recognizing a plausible task is not equivalent to establishing that the requester is authorized to initiate it.

6.2 Weak Verification Practices

Replying to a suspicious message or calling a number supplied within it may leave the user inside the attacker's communication channel. Caller identification, polished formatting, and correct grammar are also insufficient verification criteria. Independent confirmation should use contact details or authenticated services established before the suspicious interaction.

Voice and video add challenges because sensory familiarity can be mistaken for evidence of identity. Research on synthetic media and audio detection supports caution about unaided judgments of authenticity (Chesney & Citron, 2019; Warren et al., 2024). Procedures should therefore remain effective even when a message appears or sounds convincing.

6.3 Organizational Pressure and Verification Culture

Organizational routines and authority relationships can constrain verification (Heartfield & Loukas, 2015; Krombholz et al., 2015). Employees may hesitate to delay a senior leader's request or report an apparent error. Policies that reward speed without protecting legitimate challenges can create tension between operational expectations and security requirements.

Organizations should make independent confirmation routine for high-risk transactions, account recovery, sensitive data release, and privileged access. Trusted contact directories, dual approval, accessible reporting, and clear escalation routes can reduce dependence on individual discretion. Verification should establish authorization for the requested action as well as the requester's identity.

7. Human-centered Defenses

7.1 Limits of Awareness Training

Awareness training can explain manipulation tactics and verification procedures, but it cannot ensure accurate judgments in every context. Advice centered on spelling errors or unprofessional presentation is inadequate when deceptive content is fluent and plausible. Training should teach the required response to uncertainty, not imply that users can reliably identify all fabricated content.

Practical exercises should rehearse independent confirmation, payment approval, account recovery, and reporting under realistic time constraints. Evaluation should consider whether participants complete the correct verification process, not only whether they label a message suspicious. Any claimed benefit should be tested against the workload and constraints of the intended setting.

7.2 Warnings and Decision Support

Warnings should identify the relevant risk and provide a feasible next action. Repeated, vague alerts may compete with other demands on attention. Risk-sensitive presentation and clear escalation options are preferable to undifferentiated demands for vigilance, although their effects should be evaluated empirically.

Automated decision support introduces its own reliance problems: users may accept unflagged content as safe or disregard useful alerts. Trust calibration principles support communicating system limitations and uncertainty (Lee & See, 2004). AI-mediated communication also raises questions about how users interpret machine involvement and communicative intent (Hancock et al., 2020). Tools should support verification rather than replace evidence with unsupported confidence.

7.3 Organizational Verification Culture

Verification culture concerns whether organizational expectations, resources, and incentives support appropriate checking. Employees need authority to pause consequential requests and access to reliable confirmation channels. Leaders should model compliance with these procedures, including when their own requests are challenged.

Practical safeguards include dual authorization, maintained contact directories, independent confirmation, and non-punitive reporting. These measures address the organizational routines and trust relationships social engineering exploits (Heartfield & Loukas, 2015; Krombholz et al., 2015). Their implementation should account for staffing, accessibility, and the time required to resolve uncertainty.

A coherent defense integrates training, usable warnings, technical controls, and organizational procedures. Assess its effectiveness by measuring verification completion, unauthorized actions prevented, false alarms, delays, and user workload. This broader assessment can reveal whether a control improves security or simply transfers additional work to users.

8. Research Gaps and Future Research Agenda

The selected literature provides a foundation for studying trust and verification, but it does not establish a comprehensive account of adaptive AI-enabled impersonation. The priorities below are proposed directions arising from the synthesis, not gaps established through exhaustive searching. Figure 2 summarizes the agenda; Table 2 links each domain to an empirical question and design implication.

First, research should measure calibration across text, cloned voice, video, and interactive agents. Confidence ratings should be compared with accuracy and actual verification behavior. Experiments should distinguish perceived realism from familiarity, authority, and task relevance to avoid attributing every observed effect to AI generation.

Second, organizational studies should examine verification during routine work. Field studies and realistic simulations can assess how workload, hierarchy, reporting norms, and channel accessibility shape verification. Outcomes should include unauthorized compliance, successful verification, reporting latency, and disruption to legitimate tasks.

Third, interactive deception warrants examination as an emerging research problem. Agents capable of maintaining conversations may adapt requests over successive exchanges. Studies should test resistance, escalation, disengagement, and recovery rather than assume that findings from single-message phishing generalize to extended interaction.

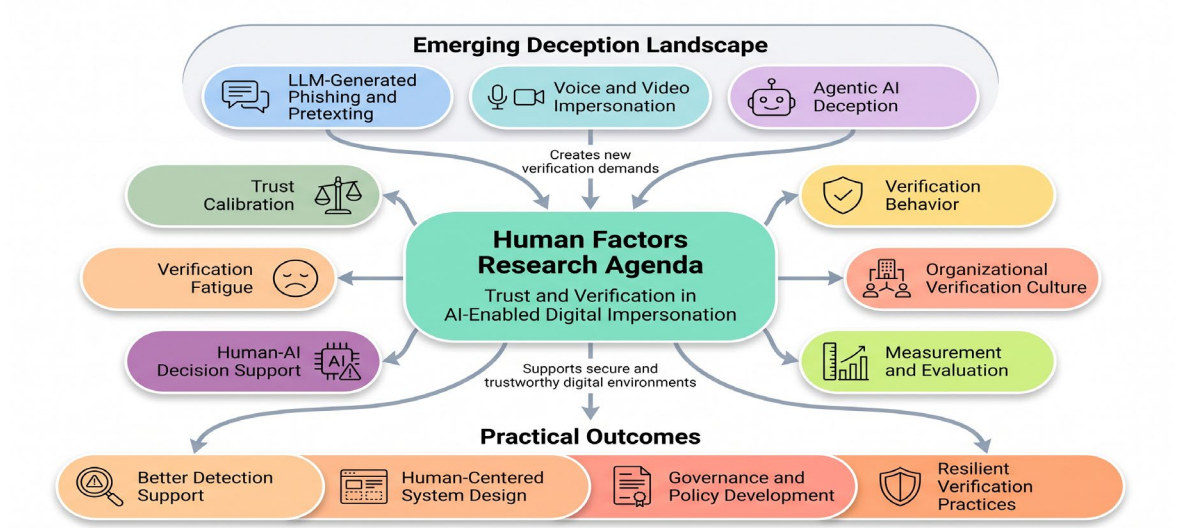

Figure 2: Forward-looking research agenda for trust and verification in AI-enabled digital impersonation

Table 2 summarizes priorities for evaluating both user behavior and the conditions that support reliable verification.

Table 2: Forward-Looking Research Agenda for Trust and Verification

Domain	Research Gap	Proposed Research Focus	Design Implication
LLM-generated phishing and pretexting	Limited evidence on how fluent AI-generated text changes susceptibility	Compare human responses to traditional phishing and LLM-generated phishing across work roles	Update training beyond grammar-based warning signs
Voice and video impersonation	Insufficient understanding of how familiar audio and visual cues influence trust	Study user confidence, uncertainty and verification behavior across cloned voice and deepfake conditions	Design verification methods that do not rely only on sensory familiarity
Agentic AI deception	Limited research on interactive AI agents that adapt persuasion over time	Examine multi-turn deception, escalation, resistance and disengagement strategies	Build safeguards for conversational verification and agent identity disclosure
Human-AI trust calibration	Limited evidence on how users trust or ignore AI-assisted detection tools	Test warning explanations, confidence indicators and user override behavior	Design decision-support tools that communicate uncertainty clearly
Organizational verification culture	Limited evidence linking workplace culture to verification success	Study reporting norms, authority pressure, workload and approval procedures	Make verification routine, safe and supported by governance
Verification fatigue	Lack of models explaining cumulative verification burden	Measure how repeated alerts and checks affect attention, compliance and decision quality	Reduce unnecessary alerts and reserve human attention for high-risk decisions

The agenda treats verification as a sociotechnical process. Governance resources, including the AI Risk Management Framework (National Institute of Standards and Technology, 2023), provide a broader basis for risk management. They should inform explicit ownership, evaluation, and accountability for verification procedures rather than be treated as evidence that a particular intervention is effective.

9. Practical Implications

Practitioners should distinguish content detection from identity and authorization checks. A realistic message can be fraudulent, and a detection score cannot independently establish permission to act. Warnings should explain what evidence is missing and direct users toward an independent confirmation process.

Organizations should embed verification in consequential workflows, including payment changes, sensitive data release, account recovery, and privileged access. Clear responsibility, protected escalation, and usable confirmation channels are necessary conditions for implementation. Monitoring should examine both prevented harm and the operational burden introduced by controls.

For policymakers, the review identifies questions concerning synthetic identity, transparency, and responsibility for verification failures. Policy analysis should distinguish general governance principles from enforceable requirements and demonstrated effectiveness. The practical objective is to distribute responsibility across system providers, organizational leaders, and users according to their capabilities and decision authority.

10. Conclusion

Digital impersonation exploits the relationship between apparent credibility and decisions made under uncertainty. Synthetic media and generated communication can reproduce familiar cues, but realism alone does not establish identity or authorization. This review integrates trust calibration, social influence, verification practices, and organizational design to explain why technical detection must be complemented by human-centered safeguards.

The synthesis supports treating verification as an organizational capability rather than an individual test of vigilance. Effective procedures must be usable, proportionate, and supported by workplace norms. Training and warnings are most defensible when they help users obtain independent evidence and complete appropriate checks.

The review's selective scope limits claims about completeness, comparative effectiveness, and causal relationships. Future research should test the proposed model across modalities and work settings, measure cumulative verification demands, and examine interactive deception. These steps would strengthen the empirical basis for systems that support appropriate trust and reliable action.

Ethics Declaration: This structured narrative review draws on published literature and publicly available materials. No primary data were collected from human participants. Accordingly, institutional review board approval was not required for the work reported here.

AI Declaration: Artificial intelligence tools assisted with literature synthesis, drafting, and editing. The authors developed and validated the conceptual framing, source selection, interpretation, and final scholarly decisions. AI tools did not independently conduct the review or determine its conclusions. The authors retain responsibility for the accuracy, integrity, and originality of the manuscript.

References

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory, 6*(3), 203–242. https://doi.org/10.1111/j.1468-2885.1996.tb00127.x

Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1820. https://doi.org/10.15779/Z38RV0D15J

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*, 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication, 25*(1), 89–100. https://doi.org/10.1093/jcmc/zmz022

Heartfield, R., & Loukas, G. (2015). A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks. *ACM Computing Surveys, 48*(3), Article 37. https://doi.org/10.1145/2835375

Krombholz, K., Hobel, H., Huber, M., & Weippl, E. (2015). Advanced social engineering attacks. *Journal of Information Security and Applications, 22*, 113–122. https://doi.org/10.1016/j.jisa.2014.09.005

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine, 18*(3), Article e1003583. https://doi.org/10.1371/journal.pmed.1003583

Siddiqi, M. A., Pak, W., & Siddiqi, M. A. (2022). A study on the psychology of social engineering-based cyberattacks and existing countermeasures. *Applied Sciences, 12*(12), Article 6042.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society, 6*(1). https://doi.org/10.1177/2056305120903408

Vishwanath, A., Herath, T., Chen, R., Wang, J., & Rao, H. R. (2011). Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems, 51*(3), 576–586. https://doi.org/10.1016/j.dss.2011.03.002

Warren, K., Tucker, T., Crowder, A., Olszewski, D., Lu, A., Fedele, C., Pasternak, M., Layton, S., Butler, K., Gates, C., & Traynor, P. (2024). "Better be computer or I'm dumb": A large-scale evaluation of humans as audio deepfake detectors. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*. Association for Computing Machinery. https://doi.org/10.1145/3658644.3670325

Workman, M. (2008). Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security. *Journal of the American Society for Information Science and Technology, 59*(4), 662–674. https://doi.org/10.1002/asi.20779