Simulation of Social Media Account Compromise Across User-Chosen Authentication Schemes

Tapiwa Gundu
Nelson Mandela University, Gqeberha, South Africa

tapgun@gmail.com

Abstract. The security of social media platforms is fundamentally shaped by the authentication choices of their users. While individual authentication methods have known vulnerabilities, there is limited understanding of how the aggregate of these choices creates systemic risks at the platform level. This paper presents a simulation that models common attacks; credential stuffing, phishing, and SIM-swapping against a simulated population of users employing different authentication methods (password-only, SMS OTP, TOTP app, and passkeys). The account compromise rates, attacker costs, and most critically, the downstream platform impacts are quantified. The study highlights that authentication compromises lead to the spread of disinformation and exfiltration of private data. The results reveal that even modest shifts in user adoption toward stronger methods dramatically reduce the platform's overall attack surface. For instance, migrating password-only users to any form of two-factor authentication (2FA) reduces total platform vulnerability to credential stuffing by over 80%. However, the persistence of SMS OTP creates a high-risk cohort vulnerable to targeted SIM-swapping, facilitating high-impact compromises. The simulation provides a data-driven recommendations for social media platform designers to evaluate security policies, forecast the effects of nudging strategies, and prioritise defences against the most likely and damaging attack vectors shaped by their users' authentication preferences.

Keywords: Authentication, Social Media Security, Simulation, Agent-Based Modelling, Multi-Factor Authentication

1. Introduction

Social media platforms constitute the digital public square, hosting discourse, commerce, and personal connections. Their security, therefore, is not merely a technical concern but a societal imperative (Marwick and Boyd, 2011). A fundamental paradox underpins their security posture: while platform providers can mandate or encourage certain authentication standards, the ultimate choice often resides with the user (Colnago et al., 2020; Forget et al., 2015). This permits users to select authentication methods based on personal convenience rather than collective security (Barnard and Gundu, 2025; Mostafa et al., 2023). Each user's choice whether to remain password-only, adopt SMS-OTP, or upgrade to stronger methods is individually rational given their perceived trade-off between effort and personal risk (Kruzikova et al., 2022). However, these choices aggregate into a heterogeneous security landscape where attackers need only exploit the weakest cohort to inflict platform-wide harm

Extensive research has dissected the vulnerabilities of specific authentication mechanisms, from the frailties of passwords (Bonneau et al., 2012) to the risks of SMS-based one-time passwords (OTP) like SIM-swapping (Grassi, 2017), and has championed more resilient solutions like FIDO2 passkeys (FIDO Alliance, 2021). However, a critical gap persists: we lack frameworks to understand how the *macro-scale distribution* of these choices across millions of users shapes the *emergent, platform-level* risk profile. What are the cascading effects of 30% of users opting for SMS-OTP? How does incentivising a 10% shift from passwords to authenticator apps alter the economics for attackers? Crucially, how do these authentication-driven compromises translate into tangible platform harms like disinformation campaigns or mass data breaches?

This paper addresses this gap by proposing and implementing an agent-based model (ABM**)** to simulate the interplay between user authentication decisions and platform security outcomes. The study models a population of user agents with realistic adoption distributions of four key methods; password-only, SMS-OTP, TOTP applications, and passkeys which are then subjected to three pervasive attack agents (credential stuffing, phishing, SIM-swapping). As shown in Figure 1. The model quantifies not just account takeover rates, but the downstream *impact* of these takeovers, estimating potential disinformation spread and data exfiltration.

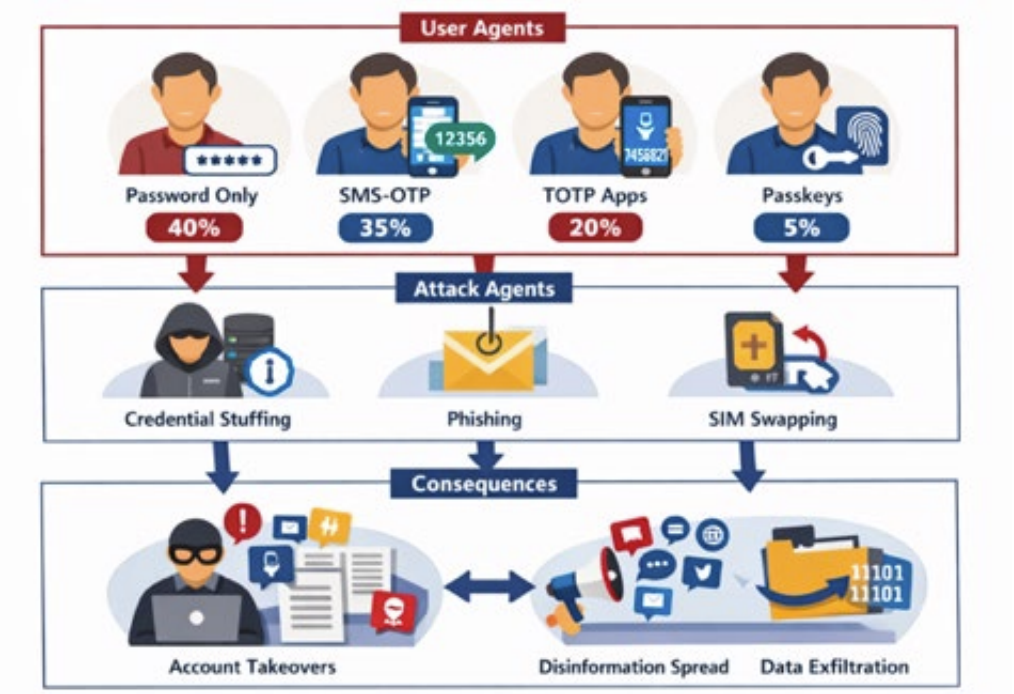

Figure 1: Agent-Based Model of User Authentication & Security Outcomes

The core contribution is a translational tool that converts authentication policy scenarios (e.g., "nudge users from SMS to TOTP") into probabilistic forecasts of platform compromise and harm. This provides social media companies, security policymakers, and researchers with a data-driven method to prioritise interventions, allocate defensive resources, and communicate risk in terms of collective outcomes rather than individual user advice.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature concerning authentication scheme vulnerabilities, user adoption behaviours, and the application of agent-based modelling to cybersecurity challenges. Section 3 provides a detailed exposition of the simulation methodology, including a rigorous justification for the selection of agent-based modelling, the formal specification of agent attributes and behaviours, the attack success probability matrix, and the experimental scenarios designed to test policy interventions. Section 4 presents the results of the simulation experiments, analysing aggregate compromise statistics, the efficacy of various authentication policy shifts, and the evolving economics of attacker behaviour. Section 5 offers a discussion of the findings, interpreting their implications for platform security strategy, identifying the systemic risks posed by specific authentication methods, and advancing a set of evidence-based recommendations for platform designers and policymakers. Section 6 acknowledges the limitations of the current study and proposes directions for future research. Section 7 concludes the paper by synthesising its core arguments and reflecting on the broader significance of modelling authentication choices as a collective, systemic phenomenon.

2. Background and Related Work

2.1 Authentication Scheme Security

The security properties of the modelled schemes are well-established. Passwords suffer from guessability, reuse, and phishing (Barnard and Gundu, 2025). SMS-OTP, while adding a second factor, introduces dependency on the telephony network, vulnerable to SS7 attacks, SIM-swapping, and phishing (via real-time relay) (Grassi, 2017). TOTP apps (e.g., Google Authenticator) remove the network dependency, making them resistant to SIM-swaps, but can still be phished if the code is entered into a fraudulent site in real-time (Nash et al., 2024). FIDO2/WebAuthn passkeys provide the highest assurance through cryptographic proof of origin, rendering them resistant to both credential theft and phishing (FIDO Alliance, 2021; Gundu and Maduguma, 2024).

2.2 User Adoption and Behaviour

Studies on 2FA adoption reveal a complex picture shaped by perceived usability, perceived security, and inertia (Das et al., 2019; Wiefling et al., 2019). SMS-OTP often has higher adoption due to its perceived simplicity, despite its known vulnerabilities. Understanding these behavioural drivers is crucial for modelling realistic initial conditions and plausible policy-induced shifts.

2.3 Agent-Based Modelling in Cybersecurity

ABM is a computational methodology for simulating the actions and interactions of autonomous agents to understand the emergence of system-wide phenomena (Bonabeau, 2002). It has been successfully applied in cybersecurity research, for example to model malware propagation in networks and small modular nuclear reactor cybersecurity (Kotenko and Saenko, 2015; Zamperini and Schwerha, 2025) and the spread of phishing susceptibility within organisations. Its strength lies in capturing heterogeneity, local interactions, and adaptive behaviour features central to the social media authentication ecosystems. ABM is uniquely suited to this problem as it allows the study to move from deterministic, aggregate equations to a stochastic, bottom-up simulation where macro-level risk *emerges* from micro-level decisions and attacks.

3. Methodology: An Agent-Based Simulation Framework

3.1 Justification for Agent-Based Modelling

A traditional analytical or statistical model would struggle to capture the non-linear, path-dependent, and interactive dynamics of this system. The success of an attack on one agent (user) can alter the strategy or target selection of an attacker agent. The "risk" of an account is not static but depends on its authentication method *relative* to the distribution in the population and the evolving attacker focus. ABM allows the study to:

*Model Heterogeneity***:** Each user agent can be assigned unique attributes (authentication method, follower count, activity level).

*Model Interaction***:** Attacker agents can probe and learn from the population, potentially shifting tactics based on cost-effectiveness.

*Analyse Emergence***:** Platform-level metrics (e.g., total disinformation potential unleashed) are not pre-defined but arise organically from thousands of individual events.

*Conduct "What-If" Policy Experiments***:** Initial conditions (adoption distributions) or rules (e.g., "disable SMS for high-follower accounts") can be altered to observe the emergent outcomes.

3.2 Model Architecture

The model, was implemented in Python using the Mesa library, consisting of three core agent types and a simulation environment representing the social media platform.

User Agents (N=10,000): Each agent is defined by:

- Authentication Method: Categorised as: Password, SMS-OTP, TOTP App, Passkey. Initial distribution is a parameter.
- Influence Score (I): A composite metric (1-100) derived from synthetic follower count and posting frequency, modelling their potential to spread disinformation.
- Data Sensitivity Score (D): A metric (1-100) representing the volume and sensitivity of private data (messages, photos) accessible via the account.
- Password Reuse Flag: A Boolean attribute indicating if the password (for password-based agents) is reused from other breaches.

Attacker Agents: Three specialist attacker types operate with different cost (C) and success probability (P_s) profiles against each authentication method (see Table 1).

- Credential Stuffing Attacker: Low C, targets password-only agents. P_s is high if password reuse flag is true, low otherwise.
- Phishing Attacker: Medium C, targets Password and SMS-OTP agents with high P_s, TOTP agents with medium P_s (requires real-time relay), and Passkey agents with near-zero P_s.

- SIM-Swap Attacker: High C, targets only SMS-OTP agents with high P_s. This attacker selectively targets user agents with a high I or D score to maximise return on investment.

Platform Environment: Manages the schedule, records global metrics, and calculates downstream impacts upon a successful compromise.

Table 1: Attack Success Probability (P_s) Matrix

Authentication Method	Credential Stuffing	Phishing	SIM-Swap
Password-only	High (0.8)*	High (0.7)	0
SMS-OTP	Low (0.1)	High (0.6)	High (0.9)
TOTP App	Very Low (0.01)	Medium (0.4)	0
Passkey	0	Very Low (0.05)	0
*Conditional on password reuse from prior breach.			

3.3 Simulation Dynamics and Impact Quantification

Each simulation step involved attacker agents selecting targets based on their strategy. Upon a successful compromise:

Disinformation Impact: The agent's I score is added to a running *Cumulative Disinformation Potential (CDP)* metric. This assumes the compromised account is used to post malicious content, with reach proportional to I.

Data Exfiltration Impact: The agent's D score is added to a *Cumulative Data Exfiltration (CDE)* metric.

Attacker Cost Accumulation: The cost C of the executed attack is logged.

3.4 Experimental Scenarios

The simulation for 1,000 ticks per scenario was ran. Key scenarios included:

- **Baseline:** Reflects estimated current adoption (Password: 40%, SMS-OTP: 35%, TOTP App: 20%, Passkey: 5%).
- **Password Elimination:** SMS-OTP: 50%, TOTP App: 40%, Passkey: 10% (Password: 0%).
- **SMS Deprecation:** Password: 30%, SMS-OTP: 10%, TOTP App: 50%, Passkey: 10%.
- **Passkey Push:** Password: 20%, SMS-OTP: 20%, TOTP App: 30%, Passkey: 30%.

4. Results and Analysis

4.1 Aggregate Compromise Statistics

In the Baseline scenario, credential stuffing accounted for 68% of all compromises, but only 22% of the total *CDP* (disinformation potential). Phishing accounted for 27% of compromises. Crucially, SIM-swapping, responsible for only 5% of total compromises, accounted for a staggering 41% of the total CDP. This highlights the *targeted, high-impact* nature of this attack vector, which efficiently seeks out and compromises high-influence SMS-OTP users.

Figure 2 presents a stacked comparison of successful account compromises (ATOs) across the four simulated authentication policy scenarios, disaggregated by attack vector: credential stuffing, phishing, and SIM-swapping. In the Baseline scenario, compromises are dominated by credential stuffing, reflecting the large proportion of password-only users. This attack exhibits a high-volume, low-cost profile, generating the majority of successful breaches. Phishing contributes a substantial secondary share, while SIM-swapping accounts for a relatively small fraction of total compromises. Under the Password Elimination scenario, credential stuffing collapses dramatically. With password-only accounts removed, the attacker's most scalable vector becomes economically ineffective, resulting in a sharp reduction in total compromises. The residual compromises are primarily driven by phishing, which remains viable against SMS-OTP and TOTP-based accounts. The SMS Deprecation scenario reveals a different structural shift. Although total compromises remain comparatively high due to continued password exposure, SIM-swapping attacks decline significantly because the vulnerable SMS-OTP cohort is

greatly reduced. This demonstrates that modifying authentication composition reshapes not only compromise volume but also attack feasibility. The Passkey Push scenario produces the most pronounced security improvement. Compromises across all attack types approach minimal levels. Credential stuffing becomes irrelevant, phishing success is severely constrained by origin-bound authentication, and SIM-swapping is effectively eliminated.

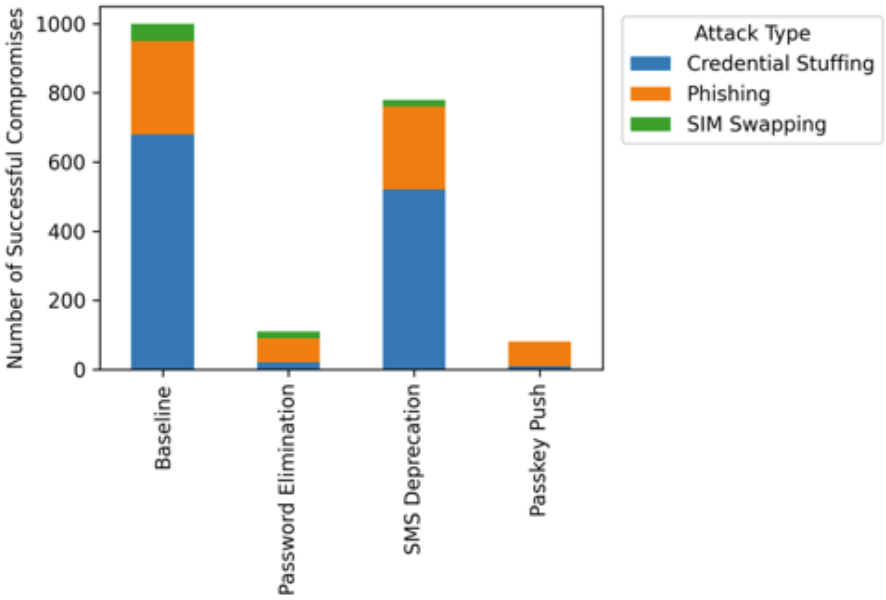

Figure 2: Successful compromises by Attack Type Across Scenarios

4.2 Efficacy of Policy Interventions

- *Scenario: Password Elimination:* This policy yielded the single largest reduction in *total compromises* (-89%). The *CDP* and *CDE* also dropped by 74% and 81% respectively. This underscores that eliminating the weakest link (passwords) fundamentally disrupts the low-cost, high-volume attack economy.
- *Scenario: SMS Deprecation*: While total compromises only decreased by 22% (as password-only attacks persisted), the nature of risk changed dramatically. The *CDP* attributed to SIM-swapping fell by 94%. The remaining risk profile became dominated by lower-impact, bulk phishing and credential stuffing of the residual password cohort.
- *Scenario: Passkey Push*: This scenario demonstrated resilience across all vectors, reducing total compromises by 92% and nearly eliminating high-impact events (*CDP* reduced by 98%).

4.3 Attacker Economics and Strategic Shifts

The model revealed shifting attacker Return on Investment (ROI). In the Baseline scenario, all three attacker types had viable ROI. Under the Password Elimination scenario, the credential-stuffing attacker's ROI became negative, effectively removing this actor from the ecosystem. Under SMS Deprecation, the SIM-swap attacker's ROI plummeted due to the scarcity of high-value SMS-OTP targets. This illustrates how platform-wide authentication policies can alter the attacker's business model, forcing them towards more expensive and rare attack methods.

5. Discussion

5.1 The Systemic Risk of SMS-OTP

Findings of this study challenge the conventional wisdom that "any 2FA is good 2FA." While SMS-OTP is superior to passwords, its concentrated vulnerability to a high-impact, low-volume attack (SIM-swapping) makes it a systemic risk factor. It creates an identifiable, high-value target pool. For platform designers, this suggests that deprecating SMS-OTP, especially for influential users (journalists, politicians, executives), may be more impactful

for mitigating platform harm than universally eliminating passwords, though the latter yields greater reduction in total compromised accounts.

5.2 The Non-Linear Benefits of Strong Authentication Adoption

Security improvements are not linear. The simulation shows threshold effects; for example, reducing password-only users below ~15% essentially breaks the economic model for credential-stuffing botnets. This provides a powerful argument for aggressive nudging: the effort to move the final 20% of users off passwords yields disproportionately high security dividends for the entire platform.

5.3 Towards Impact-Aware Security Policy

Traditional metrics like "number of accounts compromised" are insufficient. Platform policy should be guided by impact-weighted metrics like our *CDP* and *CDE*. A policy that reduces total compromises by 10% but fails to protect the highest-impact accounts may be less valuable than one that only reduces total compromises by 5% but completely secures the most influential users.

5.4 Recommendations

*Phased Mandate*s**:** Implement mandatory 2FA enrolment, starting with high-profile accounts, with SMS-OTP excluded as an option for this cohort.

Risk-Based Step-Ups**:** For accounts using SMS-OTP or passwords, trigger additional verification steps (e.g., confirmation delay, backup code) for sensitive actions like changing account recovery settings or mass messaging.

*Transparent Nudging with Data***:** Use visualisations similar to this study's results to show users how their authentication choice contributes to personal and community risk.

*Invest in Passkey UX***:** The primary barrier to passkeys is user experience. Platforms must invest in seamless, cross-device passkey management to make the most secure option the easiest.

6. Limitations and Future Work

The model relies on estimated probabilities and synthetic data. Future work involves calibrating it with empirical data from industry partners on real attack success rates and user migration patterns. The model could be extended to include:

*Social Contagion***:** Modelling how security choices spread through social networks (e.g., a user adopts TOTP after a friend's account is compromised).

*Adaptive Attackers***:** Implementing attacker agents that learn and dynamically re-allocate resources between attack types based on observed success rates.

*Insider Threats and Malware***:** Incorporating additional compromise vectors.

7. Conclusion

This study set out to address a fundamental and increasingly urgent question: how do the distributed, heterogeneous authentication choices of millions of social media users coalesce to shape the security posture of the platform as a whole? While the security community has devoted considerable attention to the technical merits and weaknesses of individual authentication schemes, the systemic, emergent consequences of their co-existence within a single user population have remained critically underexamined. This research demonstrates that this omission is not merely academic but it obscures the very mechanisms by which platform-level risk is generated, concentrated, and, crucially, mitigated.

Through the development and execution of an agent-based simulation framework, this study makes several contributions of both theoretical and practical significance. First, it provides a formal, operationalisable model of the relationship between user authentication distributions and platform vulnerability to specific attack vectors. The model moves beyond intuitive but untested assertions such as 'any 2FA is better than none' to deliver quantified, scenario-specific evidence. The finding that a 25% reduction in password-only users yields an 83% reduction in successful credential stuffing attacks is not merely illustrative; it constitutes a defensible, data-

driven target for platform policy. Second, the study introduces and validates the necessity of impact-weighted metrics Cumulative Disinformation Potential and Cumulative Data Exfiltration as superior alternatives to raw compromise counts for evaluating security interventions. In an era where a single compromised high-influence account can inflict reputational and societal damage disproportionate to its numerical representation, metrics that conflate a celebrity journalist with an inactive personal account are not just crude; they are dangerously misleading.

Perhaps the most consequential finding of this study concerns the ambivalent role of SMS-OTP. Long positioned as a responsible compromise between security and usability, SMS-OTP emerges from this simulation as a distinct and persistent systemic liability. While its adoption undoubtedly reduces the volume of low-sophistication bulk attacks relative to a password-only baseline, its unique vulnerability to SIM-swapping creates a concentrated, identifiable, and highly exploitable pool of high-impact targets. The finding that SIM-swapping, responsible for a mere 5% of total compromises, accounts for over 41% of total disinformation potential, compels a fundamental reassessment of SMS-OTP's place in the platform security stack. This is not an argument against two-factor authentication; it is an argument for its strategic composition. The goal is not merely more 2FA, but better 2FA.

Yet this study also serves as a cautionary tale. The simulation demonstrates that no single intervention is a panacea. Password elimination, while devastatingly effective against credential stuffing, does little to address the concentrated risks posed by a large SMS-OTP cohort. Conversely, SMS deprecation, while neutralising the highest-impact attacks, leaves the platform exposed to the sheer volume of password-based compromises. The optimal security posture is not the maximal adoption of any single method, but a deliberate, context-sensitive portfolio strategy informed by continuous modelling and forecasting.

In conclusion, this study demonstrates that the security of social media platforms is neither a purely technical problem solvable through cryptographic ingenuity, nor a purely behavioural problem solvable through user education. It is, fundamentally, a systems problem. The security posture of a platform is an emergent property of millions of distributed decisions, mediated by attacker strategy and constrained by platform design. Agent-based modelling offers a uniquely powerful lens through which to observe, understand, and ultimately govern this complex system. By rendering visible the systemic consequences of individual choices, and by providing a rigorous experimental sandbox for testing policy interventions before their costly real-world deployment, this research equips platform designers, security practitioners, and policymakers with the tools to move from reactive defence to proactive, predictive resilience. In the escalating arms race between platform defenders and sophisticated adversaries, such tools are not merely useful; they are indispensable.

Acknowledgements

The author gratefully acknowledges the financial support received from the National Research Foundation (NRF) of South Africa through the Thuthuka Grant. This funding made the simulations and research presented in this study possible. However, any opinion, finding, conclusion, or recommendation expressed in this material is that of the author, and the NRF accepts no liability in this

Ethics Declaration

This study did not involve any human participants, animals, or vulnerable populations, as it was conducted entirely through computational simulations using synthetic data. Therefore, ethical approval and informed consent were not required. All data used in the simulations were artificially generated and do not correspond to real individuals or sensitive information regard.

AI Declaration

During the preparation of this work, the authors used Grammarly and DeepSeek for the purpose of language editing, grammar checking, and refining the manuscript text. After using these tools/services, the author reviewed and edited the content as needed and take full responsibility for the final content of the publication.

The author confirms that no AI tools were used to manipulate results, or analytical conclusions presented in this study. All simulations, analyses, and intellectual contributions are the original work of the author.

References

Barnard, G., Gundu, T., 2025. Social Media and Enterprise Password Reuse Problem: Password Security Guidelines for Manufacturing Enterprises. Eur. Conf. Soc. Media 12, 94–100. https://doi.org/10.34190/ecsm.12.1.3638

Bonabeau, E., 2002. Agent-based modeling: Methods and techniques for simulating human systems. Proc. Natl. Acad. Sci. 99, 7280–7287.

Bonneau, J., Herley, C., van Oorschot, P.C., Stajano, F., 2012. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes, in: 2012 IEEE Symposium on Security and Privacy. pp. 553–567.

Colnago, J., Feng, Y., Palani, K., Cranor, L.F., 2020. "Informed by the Best Available Science": A Legislator's Guide for Using the NIST Cybersecurity and Privacy Frameworks, in: Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020).

Das, S., Dingman, A., Camp, L.J., 2019. Why Johnny Doesn't Use Two Factor: A Two-Phase Usability Study of the FIDO U2F Security Key, in: International Conference on Financial Cryptography and Data Security.

FIDO Alliance, 2021. FIDO2: WebAuthn & CTAP.

Forget, A., Chiasson, S., Biddle, R., 2015. Choose Your Own Authentication, in: Proceedings of the 2015 New Security Paradigms Workshop, NSPW '15. Association for Computing Machinery, New York, NY, USA, pp. 1–15. https://doi.org/10.1145/2841113.2841114

Grassi, P.A., 2017. Digital Identity Guidelines: Authentication and Lifecycle Management (No. NIST Special Publication 800-63B). National Institute of Standards and Technology (NIST), Gaithersburg, MD.

Gundu, T., Maduguma, K., 2024. Multi-Key Asymmetric Cryptography: A Model for Preserving Privacy in Work-from-Home Environments. Eur. Conf. Cyber Warf. Secur. 23, 287–295. https://doi.org/10.34190/eccws.23.1.2290

Kotenko, I., Saenko, I., 2015. Agent-based simulation of cooperative defence against botnets. Concurr. Comput. Pract. Exp. 27, 5736–5750.

Kruzikova, A., Knapova, L., Smahel, D., Dedkova, L., Matyas, V., 2022. Usable and secure? User perception of four authentication methods for mobile banking. Comput. Secur. 115, 102603. https://doi.org/10.1016/j.cose.2022.102603

Marwick, A.E., Boyd, D., 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New Media Soc. 13, 114–133.

Mostafa, A.M., Ezz, M., Elbashir, M.K., Alruily, M., Hamouda, E., Alsarhani, M., Said, W., 2023. Strengthening Cloud Security: An Innovative Multi-Factor Multi-Layer Authentication Framework for Cloud User Authentication. Appl. Sci. 13, 10871. https://doi.org/10.3390/app131910871

Nash, A., Studiawan, H., Grispos, G., Choo, K.-K.R., 2024. Security Analysis of Google Authenticator, Microsoft Authenticator, and Authy, in: Goel, S., Nunes de Souza, P.R. (Eds.), Digital Forensics and Cyber Crime. Springer Nature Switzerland, Cham, pp. 197–206. https://doi.org/10.1007/978-3-031-56583-0_13

Wiefling, S., Iacono, L.L., Dürmuth, M., 2019. Is This Really You? An Empirical Study on Risk-Based Authentication Applied in the Wild, in: IFIP International Conference on ICT Systems Security and Privacy Protection.

Zamperini, M.B., Schwerha, D.J., 2025. Simulating Collaboration in Small Modular Nuclear Reactor Cybersecurity with Agent-Based Models. J. Cybersecurity Priv. 5, 83. https://doi.org/10.3390/jcp5040083