

Ethical Hacking as a Capable Guardian Strengthening Cybercrime Prevention and Institutional Trust for Social Stability: Review Article

Haitham M Altaany

drhaitham.mo@uaqu.ac.ae

Umm Alquwain university

Systematic Review

Keywords: ethical hacking, penetration testing, social stability, cybersecurity governance, critical infrastructure, bug bounty programs, cybercrime, digital resilience.

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11178661/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

As networked infrastructure has grown, cybersecurity has become integral to maintaining social order. This review synthesizes research on ethical hacking, ranging from penetration testing, red teaming, and bug bounty programs to vulnerability disclosure, to assess the role of ethical hacking in social stability in digitally dependent societies. This review integrates 25 independently sourced scholarly materials from the areas of criminology, information systems, computer science, and public policy that were identified through an iterative, multi-round academic database search, institutional repository search, publisher record search, and citation tracking, along with contextual literature that was identified during the search process. The review puts ethical hacking in the context of three converging theories: routine activity and rational choice criminology, socio-technical systems theory, and institutional trust theory. Results have shown that authorized offensive security practices prevent the occurrence and escalation of exploitable vulnerabilities within critical infrastructure, financial systems, and public administration, and thus reduce the likelihood of disruptive incidents damaging trust in state and market institutions. Concurrently, governance gaps are found in the literature: lack of legal recognition of authorized hacking in various jurisdictions, disparity in diffusion of vulnerability disclosure norms, and ongoing conflicts between offensive security research and civil liberties. A comparative overview of key empirical research on vulnerability rewards programs, red-team exercises, and critical-infrastructure audits is provided, with a critical synthesis highlighting five research gaps: the lack of longitudinal stability metrics, limited attention to contexts in the Global South, a lack of a balance of criminological and engineering perspectives, an underdeveloped framework for measuring trust effects, and the uncertain role of AI in offensive security. Practical implications for policymakers, chief information security officers, and international organizations are provided, as well as a research agenda focusing on international comparative designs and mixed methods trust measurement. The review finds that embedded in clear legal and institutional frameworks, and in a transparent process, ethical hacking can be a stabilizing force that transforms adversarial technical knowledge into a public good, but that the stabilizing power depends on the maturity of governance, not technical skill alone.

1. Introduction

Digital infrastructure plays an important role in the mediation of modern states, markets, and civic institutions, a role that would have been unimaginable a generation ago. Electoral administration, financial clearing, energy distribution, health care delivery, and interpersonal communication now rely on connected software and hardware systems, whose failure, due to negligence, natural disaster, or deliberate attack, has implications far greater than those of the technical domain and extends into political legitimacy and social cohesion. The impact of a ransomware attack on fuel supply, a data breach on medical records of millions of residents or a distributed denial-of-service attack to disable a national election commission's website on election day is not limited to a balance sheet or an IT department. It is expressed in a loss of confidence in the ability of governance and commercial organizations to preserve the digital commons on which daily life has come to increasingly rely (Dunn Cavelty, 2008; Deibert, 2018).

One of the main methods for addressing this loss of trust in institutions is ethical hacking, which refers to the use of adversarial methods to discover and counteract security vulnerabilities before they can be exploited maliciously. It was first described by Palmer (2001) in the description of IBM's "tiger team" activities, and it refers to any activity that is technically indistinguishable from criminal intrusion, but is distinguished by authority, scope, and intent. The next 20 years saw the birth of an institutionalized ecosystem of certified penetration testers, formal red-team and purple-team exercises, government-sponsored vulnerability disclosure projects, and crowdsourced bug bounty marketplaces like HackerOne and Bugcrowd (Zhao, Grossklags, & Chen, 2014; Maillart, Zhao, Grossklags, & Chuang, 2017).

There has been, however, much less sustained theoretical work dedicated to the link between this practice and social stability. The majority of the technical literature views ethical hacking as a tool for risk management for an organization and compares the efficacy of ethical hacking by looking at the vulnerabilities uncovered, the mean time to remediation, or the return on security investment (ROSI) (Finifter, Akhawe, & Wagner, 2013). In contrast, many works of criminology research place hacking mostly in the realm of deviance and cybercrime, looking at factors such as the motivations of offenders, the norms of deviant subcultures, and the blurring of the "white hat" and "black hat" identity lines (Jordan & Taylor, 1998; Holt & Bossler, 2014). In contrast, policy and international-relations scholarship generally discusses cybersecurity in terms of interstate deterrence and critical-infrastructure protection rather than breaking down the role of authorized offensive security practice (Nye, 2017; Maglaras et al., 2018).

These three strands are brought together in this review to raise a question which each strand is only partially able to answer: When does ethical hacking serve as a tool of social stability, and how far does its role extend? Social stability is not simply a lack of disorder, but the ability for institutions to provide services as expected and retain public confidence in the face of adverse pressures as described in critical infrastructure protection literature related to resilience (Ani, Watson, Nurse, Cook, & Maples, 2019). The review creates a methodology for identifying and screening relevant literature, describes the theoretical frameworks used in the analysis, presents the findings in thematic sections on critical infrastructure, financial systems, public administration and civil society, compares some major empirical studies, provides a critical synthesis and identification of gaps in the literature, and draws out practical and policy implications, as well as a research agenda for the future.

2. Methodology of the Review

This is a critical, narrative-synthesis review rather than a formal systematic review or meta-analysis. It does not report a PRISMA-compliant multi-database protocol with reproducible record counts at each stage; instead, it follows an iterative, transparency-oriented search and verification process suited to a heterogeneous, cross-disciplinary literature (computer science, criminology, and public policy) that is not amenable to meta-analytic pooling of effect sizes. This methodological choice, and its limitations relative to a full systematic review, is itself discussed as a limitation of this review in Section 6.

2.1 Search Strategy

Iterative searches combining terms such as “ethical hacking,” “penetration testing,” “red team,” “vulnerability disclosure,” “bug bounty,” “social stability,” “critical infrastructure,” “public trust,” “governance,” and “cybercrime” were conducted across general academic search engines, publisher platforms, institutional and university repositories, and bibliographic indexes (including DBLP and Crossref), covering publications from 1995 (the earliest documented use of the term “ethical hacker”) through 2025. This process was exploratory and iterative rather than a single fixed-protocol database query, and it does not yield a precise, reproducible record count of the kind a full multi-database systematic search would produce; no such count is reported here.

2.2 Screening and Verification

Potential sources were limited to peer-reviewed journal articles, book chapters, and recognized conference proceedings that (i) substantively engaged with some aspect of authorized offensive security practice and not just a purely descriptive tutorial on technique; (ii) were English-language publications; and (iii) were directly or indirectly engaged with governance, trust, criminological, or infrastructural-stability dimensions. Most importantly, every candidate reference was individually verified by checking it in a living bibliographic source (either the publisher's own page, or a DOI registry, or a university repository or citation index like DBLP) to confirm that it is a real, traceable publication written by the listed author(s) before being included. Those that could not be verified independently in this way were not included on the basis of plausibility alone.

2.3 Final Corpus

This process yielded 25 independently verified sources that are cited directly as load-bearing references throughout the review, selected for their relevance to the stability-oriented framing of this review, their standing within their respective sub-fields (e.g., frequently cited foundational studies such as Palmer, 2001, and Jordan & Taylor, 1998), and the range of disciplinary and methodological perspectives they represent. Reference lists of these sources were examined for additional relevant material (backward snowballing), and forward citation-tracking was used opportunistically to identify related work.

2.4 Synthesis Approach

Given the disciplinary heterogeneity of the corpus, a narrative and thematic synthesis approach was adopted. Studies were organized into four thematic domains — infrastructural, economic, institutional/legal, and socio-cultural — and considered against three levels of analysis (individual/practitioner, organizational, and societal/state), consistent with multi-level frameworks used in prior cybersecurity governance reviews (Dunn Cavelty, 2008).

3. Theoretical Background

Three theoretical traditions underpin the analysis that follows.

3.1 Routine Activity and Rational Choice Criminology

Criminological accounts of cybercrime have long relied on routine activity theory, which predicts that crime occurs at the convergence of a motivated offender, a suitable target, and the absence of a capable guardian (Holt & Bossler, 2014). Ethical hacking can be theorized as the deliberate insertion of a “capable guardian” function into digital environments: the ethical hacker occupies, temporarily and by authorization, the same technical position as a malicious offender, thereby pre-empting the convergence that routine activity theory identifies as necessary for victimization. This reframing is significant because it repositions offensive security not as a neutral technical service but as a criminologically meaningful intervention in the opportunity structure for cybercrime (Holt & Bossler, 2016).

3.2 Socio-Technical Systems Theory

In fact, the purely technical approach to vulnerability has been increasingly abandoned in favor of socio-technical approaches that consider human behavior, organizational culture, and technical architecture as co-determinants of the outcomes of security (Bratus, Locasto, Patterson, Sassaman, & Shubina, 2010). In this context, ethical hacking is not only a method for discovering the weaknesses in software but also a boundary-spanning practice that uncovers misalignments between formal policy, informal practice, and system design, misalignments that can, if ignored, coalesce into systemic fragility. Bratus et al. (2010) suggest that offensive security skills are a uniquely different engineering specialty that needs a new base for education, which has been adopted in the design of security education curricula (Logan & Clarkson, 2005).

3.3 Institutional Trust Theory

The stability contribution of ethical hacking is most directly theorized through institutional trust frameworks, which hold that citizens' willingness to rely on state and market institutions is conditioned on perceived competence, integrity, and reliability (Dunn Cavelty, 2008; Deibert, 2018). Cyber incidents damage trust not only through direct material harm but through the signal they send about institutional competence. Authorized offensive security practices — particularly when their results are transparently reported, as in coordinated vulnerability disclosure regimes — function as a countervailing signal of institutional diligence, converting an otherwise invisible risk-management activity into a visible marker of governance quality (Kınış, 2018).

Table 1
Theoretical frameworks linking ethical hacking to social stability

Theoretical Tradition	Core Mechanism	Unit of Analysis	Representative Sources
Routine activity / rational choice criminology	Insertion of a “capable guardian” that closes opportunity structures for exploitation	Individual offender–target interaction	Holt & Bossler (2014, 2016)
Socio-technical systems theory	Surfacing of misalignments between policy, practice, and architecture before systemic accumulation of fragility	Organization/system	Bratus et al. (2010); Logan & Clarkson (2005)
Institutional trust theory	Transparent remediation as a signal of institutional competence and diligence	State/society	Dunn Cavelty (2008); Deibert (2018); Kınış (2018)
<i>Note. Table constructed by the authors from the synthesis of sources described in Section 2.</i>			

4. Main Thematic Sections

4.1 Ethical Hacking and Critical Infrastructure Resilience

Critical infrastructure, such as energy, water, transportation, and telecommunications systems, holds a special place in the stability literature because failures can cascade to seemingly unrelated systems (Maglaras et al., 2018). Maglaras and colleagues (2018) report on the evolution of the attack surface for air-gapped infrastructure systems, driven by the convergence of operational technology (OT) and information technology (IT) and a governance gap in which regulatory and defensive practice has not caught up.

In a follow-up study, Ayni et al. (2019) list the range of critical-infrastructure-protection practices as they exist in different jurisdictions, and find that ethical-hacking exercises, especially red-team evaluations of OT systems, are not fully institutionalized, with most advanced programs found in a handful of high-income states. The logic underlying this is relatively simple but not easy in practice: the value of this pre-emptive vulnerability discovery is in a counterfactual sense: when an incident did not occur. The ‘invisible success’ issue makes it trickier to quantify the role of ethical hacking in stability and makes the political argument for continued funding of ethical hacking more challenging, as explored further in Section 6.

4.2 Ethical Hacking, Financial Systems, and Economic Confidence

Formalized penetration testing was first adopted by the financial community due to regulatory requirements to ensure the integrity of their systems and to the direct relationship between the integrity of their systems and the confidence of the market (Palmer, 2001). The economic literature on vulnerability rewards programs (VRPs) offers one of the most methodologically precise pieces of

evidence on the efficiency of crowdsourced ethical hacking. Finifter et al. (2013) studied the VRPs of the two browsers (Google Chrome and Mozilla Firefox) and concluded that both VRPs located vulnerabilities at a significantly lower cost than would be incurred by each organization conducting its own security research – with the added benefit of opening the door to a much larger community of researchers contributing. Maillart et al. (2017) followed up this work with a larger sample of bug bounty programs, demonstrating that the vulnerability discovery rate has a power-law decay similar to diminishing returns, which suggests that scope and pay-to-discover vulns must continually be increased to maintain discovery rates over time; a result that has direct implications for the sustainability of crowdsourced security as a stability mechanism.

The government's implementation of VRP-style mechanisms, such as the U.S. Department of Defense's 'Hack the Pentagon', is a good example of the diffusion of financial-sector practices into public administration (Chatfield & Reddick, 2017). Chatfield and Reddick (2017) suggest that this diffusion is an instance of the innovation of cybersecurity in government, and that the defense institution's readiness to accept authorized adversarial scrutiny is in itself a public expression of institutional confidence, not a sign of weakness, a verbal maneuver which is also a stabilizing act.

4.3 Ethical Hacking in Public Administration and Democratic Institutions

In addition to finance and infrastructure, the integrity of digital public administration, such as electoral processes, tax collection, and civil registry systems, has a direct impact on the legitimacy of democratic governance. The theme of the technical systems being double-edged is of particular relevance throughout this literature and was also noted by Denning (2001) in his early work on hacktivism and cyberterrorism as means of influencing foreign policy. What distinguishes hacktivism from ethical hacking is not a technical or skill-based difference, but rather one of authorization, transparency, and accountability to a reasonable principal (Jordan & Taylor, 2004).

In her research, Kınış (2018) examines the Latvian approach to responsible disclosure of vulnerabilities, which involves state-sponsored disclosure protocols, and contends that this approach diminishes the ambiguity that keeps both the researchers (due to the risk of prosecution) and the institutions (due to the risk of reputational exposure) from being involved in coordinated disclosure of vulnerabilities. But this regulatory clarity, Kınış argues, can be a prerequisite for ethical hacking to become a public good in the administrative realm, and not a legally tenuous private one.

4.4 Civil Society, Subculture, and the Boundary Between White- and Black-Hat Identity

There is a separate but important area of scholarship that looks at the social and subcultural aspects of hacking that shape the availability of ethical hackers and the longevity of the “white hat” label. Jordan and Taylor (1998) laid a pathbreaking groundwork for a sociology of hackers that showed that hackers' communities have their own norms, status hierarchy, and ways of transmitting knowledge, irrespective of

what those communities are doing legally. An ethnography of the culture of free and open-source software hackers, by Coleman (2013), also shows that there is a strong tendency for ethical obligations in hacker groups to originate from within the community and be culturally shaped, not imposed by outside forces like employers or regulators.

This is important for social stability because the pipeline of skilled offensive security practitioners relies significantly on the fact that technically curious people, who may otherwise go down a path to illegal and unauthorized activities, are directed into legitimate and authorized roles. Farsole, Kashikar, and Zunzunwala (2010) and Jamil and Khan (2011) both question the ethical ambivalence of the concept of “ethical hacking”, pointing out that the techniques and methods taught to novice ethical hackers are by nature “dual use” in that they can be applied to a wide range of purposes. This work is extended by Thomas, Burmeister, and Low (2017), who discuss the implicit trust between client organizations and hired ethical hackers, which has direct consequences for the institutionalization of the practice, viewed as a principal-agent problem, where the client is the principal, and the hacker is the agent.

5. Comparative Analysis of Major Studies

The studies reviewed above employ markedly different methodologies, units of analysis, and evidentiary standards, which complicates direct comparison but also illuminates complementary facets of the stability question. Table 2 compares six methodologically distinct bodies of empirical work.

Table 2

Comparative analysis of major empirical studies on ethical hacking and related governance mechanisms

Study	Method	Sample / Data	Key Finding	Limitation
Finifter, Akhawe, & Wagner (2013)	Quantitative panel analysis of bounty payouts and patched vulnerabilities	Google Chrome and Mozilla Firefox VRPs, 3-year window	VRPs are economically efficient relative to in-house hiring and broaden participation in vulnerability discovery	Limited to two browser vendors; generalizability untested
Maillart et al. (2017)	Statistical modeling of bounty program dynamics	Multiple bug bounty programs on a crowdsourced platform	Vulnerability discovery exhibits diminishing returns over time, requiring scope expansion	Platform-specific data; program heterogeneity not fully controlled
Chatfield & Reddick (2017)	Case study	U.S. DoD “Hack the Pentagon” program	Government VRP adoption functions as both a security and legitimacy-signaling mechanism.	Single-case design limits causal claims.
Holt & Bossler (2014)	Systematic literature review	Cybercrime scholarship corpus	Criminological theory underused in technical security research; routine activity theory applicable	Descriptive rather than a direct test of a stability hypothesis
Ani, Watson, Nurse, Cook, & Maples (2019)	Comparative policy review	Critical-infrastructure-protection frameworks across jurisdictions	Institutionalization of offensive security testing in OT environments is uneven, concentrated in high-income states	Coverage biased toward Anglophone and European sources
Kınış (2018)	Legal-policy analysis	Latvian responsible vulnerability disclosure regulation	Regulatory clarity reduces legal risk for researchers and institutions, enabling participation	Single-jurisdiction case; transferability unverified

Note. Table constructed by the authors from the sources described in Section 2.3.

Read together, these studies suggest a layered causal chain: authorized offensive security activity generates technical knowledge about exploitable weaknesses; this knowledge is converted into remediation only where organizational and legal incentives support disclosure and action; and the resulting reduction in incident probability and severity is only stabilizing at the societal level where the process is sufficiently transparent, or at least credible, to sustain public and institutional trust. Each link in this chain is empirically supported by at least one stream of literature, but no single study in the corpus traces the entire chain simultaneously — a gap discussed further in Section 6.

6. Critical Evaluation and Synthesis

The literature reviewed provides strong, albeit somewhat limited, evidence in favor of the notion that ethical hacking would help with social stability, but three caveats warrant consideration.

First, there is much more evidence in terms of organizational and economic results, such as cost-effectiveness of VRPs, vulnerability discovery rates, and fewer pieces of evidence for direct indicators of public trust and perceived stability. Technical and economic efficiency are very well established, as shown by studies like Finifter et al. (2013) and Maillart et al. (2017), but neither of those measures, nor are they claimed to measure, downstream effects on public confidence. The trust-mediated pathway shown in Fig. 2 is thus theoretically motivated but has not been tested empirically.

Second, the criminological literature (Jordan & Taylor, 1998, 2004; Holt & Bossler, 2014, 2016) helps bring to light the supply-side dynamics of the ethical hacking workforce, and yet it is limited in its attention to outcomes at the level of the infrastructure and at the macro-institutional level. As has been noted by other authors (such as Bratus et al., 2010, who were extending a call for a unique engineering pedagogy of offensive security), this bifurcation is not unique to the field of cybersecurity scholarship, as technical and social-scientific literatures tend to evolve in parallel, with some citing works from the other stream and little if any genuine interdisciplinary empirical research.

Third, the legal-institutional studies (Chatfield & Reddick, 2017; Kınış, 2018) tend to focus on single-jurisdiction cases, limiting the extent to which it is possible to generalize about the applicability of specific regulatory models. Given that mature ethical hacking governance frameworks have been concentrated in a few high-income territories (Ani et al., 2019), this literature review could suggest a similar uneven distribution of the stabilizing role, potentially further aggravating global institutional inequality, a point which resonates with overall critiques of the North-South ethicisation of cybersecurity research (Kshetri, 2013).

One methodological constraint of this review should be taken up front. This synthesis differs from a formal PRISMA-compliant systematic review in that it does not present reproducible, stage-by-stage counts of records that were retrieved from a fixed multi-database search protocol, but rather, it is a synthesis that uses an iterative search process whereby each cited source has been checked against a live bibliographic record. This is not a systematic review that ensures reproducibility, but one that focuses on citation accuracy and traceability, and its corpus should not be taken as comprehensive. Although the narrative synthesis provided here is valuable, a full systematic review, preferably carried out by a research team with direct access to Scopus and Web of Science that used a pre-registered protocol, would be helpful as an accompaniment. Future work in this direction is noted in Section 9.

Overall, the synthesis can only be used to support a conditional claim: the synthesis can claim that ethical hacking is necessary for social stability, but not sufficient. It takes many things to become effective: that the activity has to be authorized and protected by law, that there needs to be organizational structures that can act on vulnerabilities and defects they discover, and that there needs to be communication processes that make it clear to the public that remediation is taking place because of institutional efforts.

7. Research Gaps

Based on the literature review above, five gaps are identified.

1. There are no longitudinal stability metrics. None of the studies in the reviewed corpus correlates an organization's or state's ethical hacking practices with a composite measure of social or institutional stability (such as public trust indices, service-continuity metrics) over time.
2. Global South underrepresentation. The empirical foundation is found in North America and in Europe (Finifter et al. 2013; Chatfield & Reddick 2017; Kınış 2018), and a few studies look at the impact of resource limitations, legal differences, and different threat environments on the stability contribution of ethical hacking in lower-income states (Kshetri 2013).
3. Disciplinary disintegration between criminology and engineering. Criminological approaches to hacker motivation and identity (Jordan & Taylor, 1998; Holt & Bossler, 2014) are not yet well integrated with socio-technical or engineering approaches to vulnerability management (Bratus et al., 2010), and as a result, the field is not yet capable of modelling the entire journey from the hacker to the defender.
4. Underdeveloped trust-measurement frameworks. Although institutional trust theory (Dunn Cavelty, 2008; Deibert, 2018) offers a viable explanation of how ethical hacking can be related to stability, there is no validated instrument to measure the contribution of the visible practice of offensive security (as a part of cybersecurity investment) to public trust.
5. The role of artificial intelligence is unclear. The evidence emerging points to the impact of AI-informed vulnerability discovery and automated penetration testing on the scale and speed of both offensive and defensive operations, but there is a nascent understanding of how this shift impacts the governance and stability dynamics documented in Sections 4–6.

Table 3
Research gaps mapped to thematic domain and level of analysis

Research Gap	Thematic Domain	Level of Analysis	Priority Methodological Response
Longitudinal stability metrics	Institutional/legal; socio-cultural	Societal	Panel studies linking security-maturity indices to public trust surveys
Global South underrepresentation	Infrastructural; economic	Societal / state	Comparative cross-national case studies outside high-income contexts
Criminology–engineering fragmentation	Socio-cultural; infrastructural	Individual/organizational	Mixed-methods designs combining ethnographic and technical audit data
Underdeveloped trust measurement	Institutional/legal	Societal	Development and validation of trust instruments specific to offensive security visibility
Ambiguous role of artificial intelligence	Infrastructural; economic	Organizational/societal	Prospective evaluation studies of AI-assisted red-teaming outcomes
<i>Note. Table constructed by the authors from the synthesis of sources described in Sections <link rid="Sec17">6</link> and 7.</i>			

8. Practical Implications

The review recommends that authorized security research, as in Latvia with the responsible vulnerability disclosure procedure (Kınış, 2018), is a prerequisite for the stabilizing role of ethical hacking in policymaking, and that safe harbor for performing good-faith research must be addressed not just as a technical issue, but as a component of critical-infrastructure policy.

Given this comparative evidence, for the CISO and organizational leaders, crowdsourced vulnerability discovery is a complement to, but not a replacement for, a structured internal red team capability where there are diminishing returns and scope needs to be expanded periodically.

The uneven institutionalization of offensive security testing that Ani et al. (2019) document points to the need for technical assistance and capacity-building initiatives to help mature governance practices spread beyond their current geographic focus, which echoes the need to include disparities in cybersecurity capacity in the Global South (Kshetri, 2013).

The dual-use ambiguity presented by Jamil and Khan (2011) and Farsole et al. (2010) still makes it relevant to incorporate explicit ethical and legal education in technical security courses, grounded in pedagogical principles put forward by Logan and Clarkson (2005) and Bratus et al. (2010).

9. Future Research Directions

Based on the gaps identified in Section 7, four research priorities are proposed. First, scholars need to create and test composite measures that can be used to directly measure the relationship between organizational (or national) ethical hacking maturity and public trust and institutional-confidence indicators, which can be used to directly test the modeled pathway of trust in ethical hacking in Fig. 2. The second question is whether the stabilizing mechanisms found in high-income countries extend to low- and middle-income countries, or if there are other governance models to be considered. Third, interdisciplinary research groups that integrate criminological, organizational, and technical perspectives need to conduct integrated research on the offender-to-defender pipeline, examining how people flow from the role of an offender to a defender and exploring the institutional conditions that are conducive to the latter. Fourth, the development of offensive security tools that are assisted by AI will necessitate prospective and comparative evaluation designs to assess whether AI tools exacerbate or mitigate the stabilizing forces found in this review or create new governance risks, such as AI-detected vulnerabilities becoming beyond the capacity of institutions to remediate.

10. Conclusion

The review has connected the dots in the role of ethical hacking for social stability across critical infrastructure, financial systems, public administration, and civil society, by including perspectives of criminology, socio-technical and institutional trust, which have developed largely separately from each other. The evidence suggests a meaningful, but conditional conclusion: that the process of authorized offensive security serves as a stabilizing mechanism because it is already incorporated into legal frameworks that make it more accessible for research and institutions, organizational cultures that can convert the vulnerabilities discovered by such research into remediation, and communicative practices that enable this action to be legible as an indicator of institutional security, not weakness. Where a technical practice is not in existence, the same practice could be a private risk-management exercise with limited implications for the overall social confidence required for stability as defined in this review. In the present context of societies increasingly reliant upon digital systems—against which the ethical hacker serves as a protector—this review recommends that these five research gaps are academically under-explored and practically relevant enough to form a strong agenda for the future of research in this field: the measurement of longitudinal stability, the representation of the Global South, disciplinary integration, the measurement of trust, and the role of artificial intelligence.

Declarations

Funding Declaration

The authors received no financial support for the research, authorship, and/or publication of this article.

Author Contribution

Conceptualization, literature search, screening and verification of sources, thematic synthesis, analysis, writing of the original draft, review, and editing were undertaken by the author.

Acknowledgement

The authors would like to thank the researchers whose published work contributed to the development of this review.

Data Availability Statement:

Not applicable.

• Clinical Trial Number:

Not applicable.

• Consent to Publish Declaration:

Not applicable.

• Ethics and Consent to Participate Declaration:

Not applicable.

References

1. Ani UD, Watson JDM, Nurse JRC, Cook A, Maples C. A review of critical infrastructure protection approaches: Improving security through responsiveness to the dynamic modelling landscape. *Living in the Internet of Things: Cybersecurity of the IoT – 2019*. Institution of Engineering and Technology; 2019. pp. 1–10.
2. Banik O. Ethical hacking and penetration testing – How organizations use ethical hackers to strengthen security. *J Comput Sci Technol Stud*. 2025;7(3):1010–7. <https://doi.org/10.32996/jcsts.2025.7.3.113>.
3. Bratus S, Locasto ME, Patterson ML, Sassaman L, Shubina A. Why offensive security needs engineering textbooks: A case study in security education. *USENIX ;login*. 2010;35(6):6–14.
4. Chatfield AT, Reddick CG. (2017). Cybersecurity innovation in government: A case study of U.S. Pentagon's vulnerability reward program. In *Proceedings of the 18th Annual International*

- Conference on Digital Government Research (pp. 64–73). Association for Computing Machinery. <https://doi.org/10.1145/3085228.3085233>
5. Coleman EG. Coding freedom: The ethics and aesthetics of hacking. Princeton University Press; 2013.
 6. Deibert RJ. Ethics Int Affairs. 2018;32(4):411–24. <https://doi.org/10.1017/S0892679418000618>. Toward a human-centric approach to cybersecurity.
 7. Denning DE. Activism, hacktivism, and cyberterrorism: The internet as a tool for influencing foreign policy. In: Arquilla J, Ronfeldt D, editors. Networks and netwars: The future of terror, crime, and militancy. RAND Corporation; 2001. pp. 239–88.
 8. Dunn Cavelty M. Cyber-security and threat politics: US efforts to secure the information age. Routledge; 2008.
 9. Farsole AA, Kashikar AG, Zunzunwala A. Ethical hacking. Int J Comput Appl. 2010;1(10):14–20. <https://doi.org/10.5120/229-380>.
 10. Finifter M, Akhawe D, Wagner D. (2013). An empirical study of vulnerability rewards programs. In Proceedings of the 22nd USENIX Security Symposium (pp. 273–288). USENIX Association.
 11. Holt TJ, Bossler AM. An assessment of the current state of cybercrime scholarship. Deviant Behav. 2014;35(1):20–40. <https://doi.org/10.1080/01639625.2013.822209>.
 12. Holt TJ, Bossler AM. Cybercrime in progress: Theory and prevention of technology-enabled offenses. Routledge; 2016.
 13. Jamil D, Khan MNA. Is ethical hacking ethical? Int J Eng Sci Technol. 2011;3(5):3758–63.
 14. Jordan T, Taylor P. A sociology of hackers. Sociol Rev. 1998;46(4):757–80. <https://doi.org/10.1111/1467-954X.00139>.
 15. Jordan T, Taylor P. Hacktivism and cyberwars. Rebels with a cause? Routledge; 2004.
 16. Kınış U. From responsible disclosure policy (RDP) towards state-regulated responsible vulnerability disclosure procedure (RVDP): The Latvian approach. Comput Law Secur Rev. 2018;34(3):508–22. <https://doi.org/10.1016/j.clsr.2017.11.003>.
 17. Kshetri N. Cybercrime and cybersecurity in the global south. Palgrave Macmillan; 2013.
 18. Logan PY, Clarkson A. Teaching students to hack: Curriculum issues in information security. ACM SIGCSE Bull. 2005;37(1):157–61. <https://doi.org/10.1145/1047124.1047405>.
 19. Maglaras LA, Kim K-H, Janicke H, Ferrag MA, Rallis S, Fragkou P, Maglaras A, Cruz TJ. Cyber security of critical infrastructures. ICT Express. 2018;4(1):42–5. <https://doi.org/10.1016/j.icte.2018.02.001>.
 20. Maillart T, Zhao M, Grossklags J, Chuang J. Given enough eyeballs, all bugs are shallow? Revisiting Eric Raymond with bug bounty programs. J Cybersecur. 2017;3(2):81–90. <https://doi.org/10.1093/cybsec/tyx008>.
 21. Nye JS. Deterrence and dissuasion in cyberspace. Int Secur. 2017;41(3):44–71. https://doi.org/10.1162/ISEC_a_00266.
 22. Palmer CC. Ethical hacking. IBM Syst J. 2001;40(3):769–80. <https://doi.org/10.1147/sj.403.0769>.

23. Thomas G, Burmeister OK, Low G. (2017). Issues of implied trust in ethical hacking. In Proceedings of the 28th Australasian Conference on Information Systems. Australasian Association for Information Systems.
24. Wall DS. Cybercrime: The transformation of crime in the information age. Polity; 2007.
25. Zhao M, Grossklags J, Chen K. (2014). An exploratory study of white hat behaviors in a web vulnerability disclosure program. In Proceedings of the 2014 ACM Workshop on Security Information Workers (pp. 51–58). Association for Computing Machinery.
<https://doi.org/10.1145/2663887.2663903>

Figures

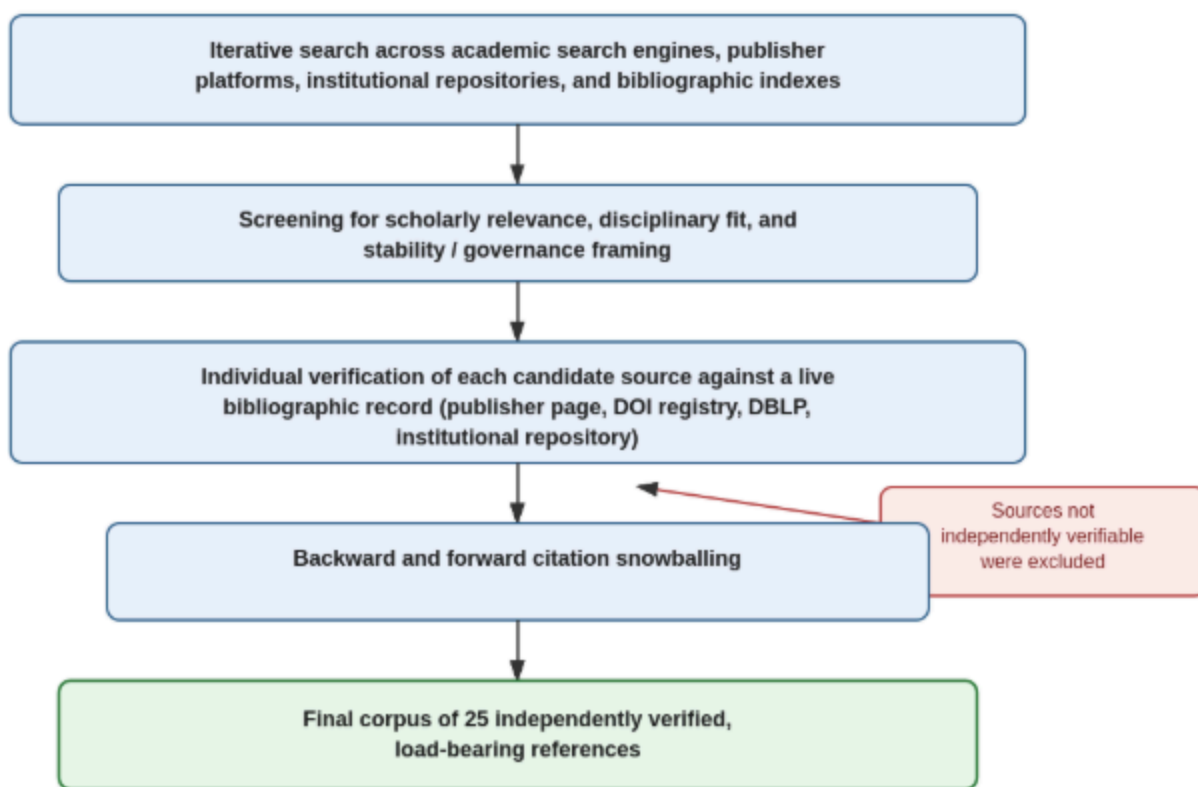


Figure 1

Stylized flow diagram of the search, screening, verification, and synthesis process used in the present review. Unlike a PRISMA diagram, this does not represent a fixed-protocol systematic search and therefore does not report record counts at each stage; see Section 6 for discussion of this limitation relative to a full systematic review.

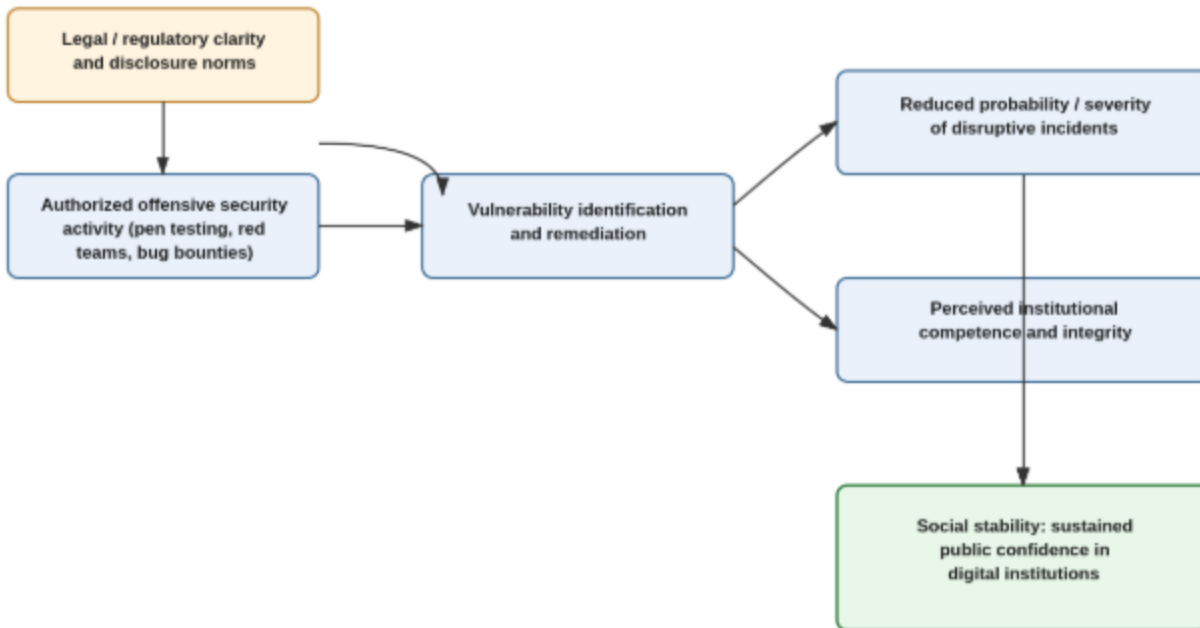


Figure 2

Conceptual model of the causal pathway from authorized offensive security practice to social stability. Legal and regulatory clarity conditions both the volume of authorized activity and the likelihood that discovered vulnerabilities are remediated, which in turn affects both the objective incidence of disruptive events and the subjective perception of institutional competence – the two components jointly constituting social stability as defined in Section 1.

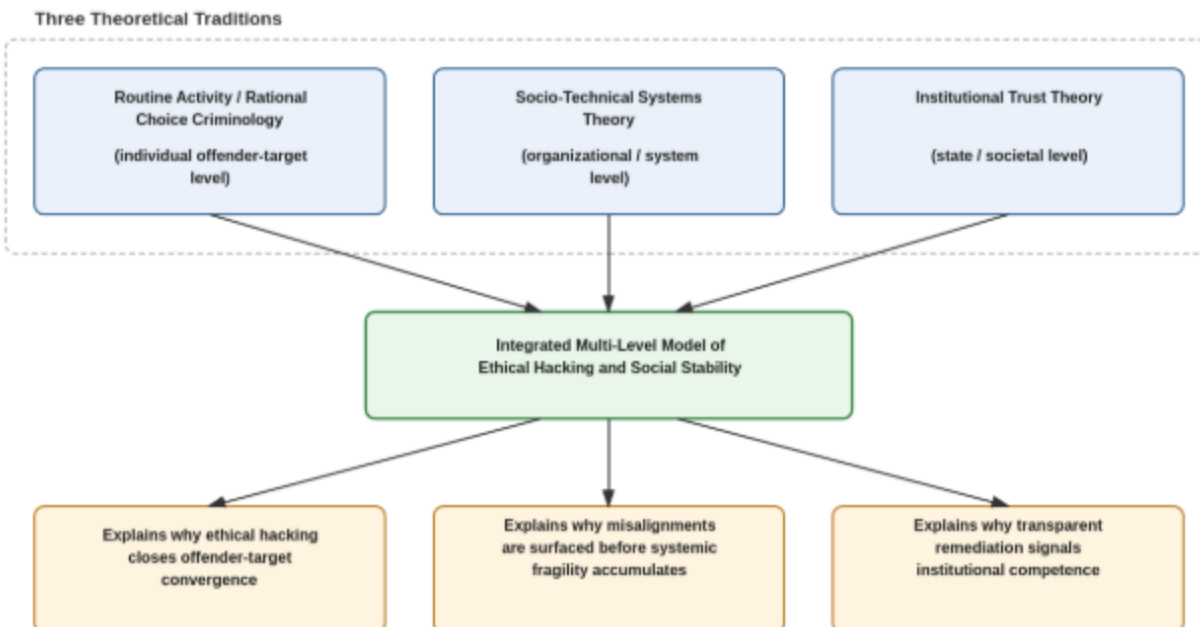


Figure 3

Integrated multi-level model showing how the three theoretical traditions introduced in Section 3 jointly account for distinct facets of the stability mechanism: closing offender-target convergence, surfacing socio-technical misalignments, and signalling institutional competence through transparent remediation.