THE UPSTREAM BLIND SPOT: PEGASUS AND SURVEILLANCE INFRASTRUCTURES AS DUAL-USE AI RISK

Anonymous authors
Paper under double-blind review

ABSTRACT

Assessments of artificial intelligence (AI) dual-use risk in security and military contexts often focus on downstream algorithms while overlooking the infrastructural conditions that enable their deployment. This paper argues that upstream surveillance infrastructures constitute a critical but underexamined locus of AI dual-use risk. Through an analysis of zero-click surveillance and a case study of Pegasus spyware, we show how automated, covert data extraction systems enable persistent, large-scale intelligence collection and condition the data environments upon which downstream analytic processes depend. We further examine how privatized surveillance markets diffuse responsibility across public and private actors, constraining oversight and accountability. The paper advances an infrastructural perspective on AI dual-use that foregrounds surveillance architectures alongside algorithmic systems.

1 INTRODUCTION

Contemporary discussions of artificial intelligence (AI) governance increasingly emphasize the dual-use risks associated with machine learning systems, particularly their incorporation into military and security practices. Existing analyses have largely concentrated on model-centric deployments, such as automated classification, decision-support, or targeting systems, while paying comparatively less attention to the infrastructural conditions that render such systems operational. This paper argues that upstream surveillance infrastructures constitute a critical but underexamined layer through which AI-enabled intelligence practices are materially enabled.

Commercial spyware platforms provide a concrete empirical setting in which to examine this gap. Pegasus spyware, developed and sold by NSO Group, is not an artificial intelligence system in itself. Rather, it operates as a covert, automated surveillance infrastructure capable of persistent and high-fidelity data extraction from personal mobile devices. Forensic investigations have demonstrated that Pegasus relies on chained zero-day and zero-click exploits, enabling device compromise without user interaction and with minimal observable traces Marczak et al. (2020; 2021); Shia (2021). These capabilities allow continuous access to communications, sensor data, and location information, substantially expanding the scope and scale of intelligence collection.

Technical analyses further document Pegasus's architectural sophistication, including privilege escalation, selective data exfiltration, encrypted command-and-control channels, and resistance to standard detection mechanisms Rudie et al. (2021); O'Cearbhaill & Marczak (2022); Chourasiya et al. (2023). Such features lower the operational barriers to large-scale surveillance and facilitate the integration of commercially developed spyware into state intelligence workflows. As prior scholarship has shown, Pegasus is embedded within a broader market for privatized espionage technologies characterized by opacity, deniability, and limited avenues for public accountability Kaster & Ensign (2023); Spens (2024); Kotliar & Carmi (2024).

This paper conceptualizes Pegasus as an upstream surveillance infrastructure that conditions downstream algorithmic analysis, rather than as an isolated instance of digital intrusion. By synthesizing forensic evidence, technical studies of zero-click exploitation, and research on mercenary spyware governance, we examine how such infrastructures reshape the epistemic and operational foundations of AI-mediated intelligence practices. We contend that assessments of AI dual-use risk remain

incomplete unless they account for the surveillance systems that supply, structure, and normalize the data environments upon which algorithmic systems depend Deibert (2022); Anstis (2024); Chan (2018).

2 From AI Systems to Intelligence Infrastructures

Analyses of artificial intelligence in security and military contexts often conceptualize AI as a discrete system, typically a model or algorithm, whose risks can be evaluated through properties such as accuracy, autonomy, or explainability. While this perspective has generated important insights, it obscures the broader infrastructural arrangements through which algorithmic systems acquire operational significance. Intelligence practices increasingly emerge from assemblages of data collection, automation, storage, and analysis, in which upstream surveillance infrastructures provide persistent high-volume, high-resolution data, enabling behavioral profiling, network inference, and prioritization by expanding the feasible domain of downstream analytic capacity.

Pegasus illustrates how such infrastructures restructure intelligence workflows. Forensic studies demonstrate that Pegasus enables sustained access to a device's communications, sensors, and metadata through chained exploits and privileged execution, while selectively exfiltrating data to external command-and-control servers Rudie et al. (2021); Shia (2021). This architecture effectively transforms personal mobile devices into continuously accessible sensors embedded within everyday social environments. The intelligence value of such systems does not derive from any single analytic technique, but from the accumulation, correlation, and subsequent processing of heterogeneous data streams.

Importantly, these infrastructural dynamics complicate conventional approaches to accountability and governance. Because upstream surveillance tools are often developed and operated by private vendors, their integration into state intelligence practices introduces layers of institutional separation and deniability. Prior research on mercenary spyware markets highlights how privatization diffuses responsibility across corporate, legal, and national boundaries, limiting oversight while expanding access to advanced surveillance capabilities Kaster & Ensign (2023); Spens (2024); Kotliar & Carmi (2024). In this context, algorithmic analysis may be formally downstream, yet substantively enabled by surveillance infrastructures that remain opaque to both public scrutiny and regulatory intervention.

Conceptualizing AI as embedded within intelligence infrastructures rather than as a standalone system therefore shifts the locus of risk. It draws attention to how automation at the level of data capture conditions the deployment, scale, and perceived legitimacy of subsequent analytic processes. Understanding AI-mediated intelligence practices thus requires examining not only models and decision-support tools, but also the upstream surveillance systems that operationalize them.

3 Pegasus as an Upstream Intelligence Infrastructure

Pegasus spyware exemplifies how upstream surveillance infrastructures enable AI-mediated intelligence practices by automating device compromise, persistence, and selective data extraction at scale. Forensic and technical analyses consistently characterize Pegasus as a modular platform that exploits chained zero-day and zero-click vulnerabilities to obtain privileged access to mobile operating systems without user interaction Marczak et al. (2020; 2021); Rudie et al. (2021). By eliminating user behavior as a gating factor, Pegasus embeds surveillance directly within routine device operation, enabling continuous and covert data capture.

Following exploitation, Pegasus establishes system-level persistence that permits sustained access to communications content, call records, contact lists, location data, and device sensors Rudie et al. (2021); Chourasiya et al. (2023). Encrypted command-and-control channels enable remote tasking and data exfiltration, while selective collection mechanisms allow operators to specify which categories of data are extracted Chourasiya et al. (2023). This selectivity is operationally significant: it reduces bandwidth costs, limits observable artifacts, and supports the structured assembly of heterogeneous datasets rather than indiscriminate bulk collection. Across forensic investigations, resistance to detection and attribution emerges as a defining property of Pegasus deployments. Exploit artifacts are frequently transient, partially overwritten, or deliberately removed, complicating

retrospective identification Shia (2021); O'Cearbhaill & Marczak (2022). In multiple documented cases, attribution relied on subtle residual indicators, such as anomalous database entries or incomplete cleanup of exploit-related files, rather than persistent malware signatures Marczak et al. (2020; 2021). These characteristics enable prolonged surveillance without triggering user awareness or conventional security alerts.

From an analytic perspective, the intelligence significance of Pegasus lies in the data environments it produces. Persistently collected communications, metadata, and location traces correspond to standard inputs for contemporary AI-based analytic systems, including natural language processing for communications analysis and graph-based models for inferring relational networks and behavioral patterns. Pegasus does not perform such inference itself; rather, its automation and persistence materially reduce the cost and uncertainty of assembling the data required for downstream algorithmic analysis.

Comparative forensic analyses across iOS and Android platforms indicate that, despite architectural differences, Pegasus achieves consistent outcomes: persistent access, cross-application visibility, and covert data extraction across device ecosystems Caruso (2024). As emphasized in comprehensive technical assessments, Pegasus functions as a data acquisition layer whose outputs are intended to be stored, correlated, and interpreted by external systems and analysts to support profiling, tracking, and investigative prioritization Kareem (2024). In this sense, Pegasus illustrates how upstream surveillance infrastructures condition the feasibility and scale of AI-mediated intelligence practices prior to model deployment.

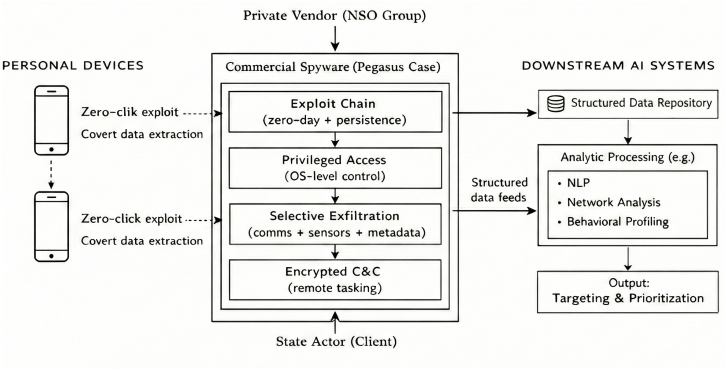

Figure 1: Pegasus as upstream surveillance infrastructure supplying structured data that can support downstream analytic processing (e.g., NLP), thereby pre-conditioning targeting and prioritization workflows.

4 Institutional Embedding and Privatized Surveillance Markets

The operation of Pegasus as an upstream intelligence infrastructure is inseparable from the institutional and market arrangements through which commercial surveillance technologies are produced, sold, and integrated into state security practices. Prior scholarship situates Pegasus within a broader ecosystem of privatized espionage, in which capabilities historically developed and maintained by state intelligence agencies are increasingly supplied by private vendors Kaster & Ensign (2023); Spens (2024). This shift reshapes how surveillance capacities are acquired, legitimized, and governed, with direct implications for downstream analytic use. Commercial spyware vendors position themselves as intermediaries between specialized technical expertise and state demand for covert surveillance, and analyses of NSO Group's business practices indicate that Pegasus is marketed exclusively to government clients and framed as a lawful instrument for counterterrorism or serious crime investigations Kaster & Ensign (2023); Kotliar & Carmi (2024). Such framing legitimizes intrusive surveillance as an extension of sovereign authority while displacing operational responsibility across contractual and organizational boundaries.

Privatization also introduces structural opacity that complicates accountability, since development, maintenance, and technical support are distributed across corporate and governmental actors. Spens characterizes this configuration as a "spyware industrial complex," in which secrecy, transnational

contracting, and competitive market dynamics limit the effectiveness of traditional oversight mechanisms Spens (2024). These dynamics reinforce the infrastructural role of systems like Pegasus: states externalize surveillance capabilities to reduce internal costs while gaining access to continuously updated exploit chains and technical support Woodhams (2021), and vendors benefit from long-term contractual relationships that incentivize refinement of covert access techniques. This normalization stabilizes the data conditions required for downstream analytic processing by ensuring steady supplies of high-resolution communications, metadata, and sensor data, thereby preconditioning AI-mediated intelligence deployment and shifting dual-use risk to stages preceding model development, training, or deployment. Section 5 examines how this upstream shift complicates prevailing approaches to accountability and governance in AI-mediated intelligence practices.

5 Discussion: Implications for AI Dual-Use and Accountability

The analysis in Sections 2–4 indicates that AI-related risk in security and military contexts cannot be adequately assessed by examining models or decision-support systems in isolation. Upstream surveillance infrastructures such as Pegasus shape the feasibility, scale, and effects of AI-mediated intelligence practices by stabilizing continuous, high-resolution data capture prior to analytic processing. As shown in Section 3, these infrastructures automate compromise and persistence, while Section 4 demonstrates how institutional and market arrangements normalize their use, relocating dual-use risk to stages that precede model development, training, or deployment. This upstream shift complicates governance approaches focused on explainability, bias, or human-in-the-loop oversight, since constraints applied at the level of analytic models may have limited effect when covert data acquisition escapes meaningful review. Prior work on mercenary spyware emphasizes that such systems often exist in regulatory grey zones, frustrating oversight while enabling expansive intelligence practices Deibert (2022); Anstis (2024).

The infrastructural perspective also clarifies epistemic asymmetries: individuals subject to surveillance lack visibility into how data are captured, retained, and later analyzed, while institutions retain discretion over downstream interpretation and use. Scholarship on algorithmic accountability cautions that transparency-based governance is insufficient under conditions of technical opacity and institutional fragmentation Ananny & Crawford (2018); Stefanija (2023). Finally, the privatized organization of surveillance markets intensifies dual-use risk by diffusing responsibility across public and private actors. As discussed in Section 4, commercial vendors provide turnkey surveillance capabilities while maintaining contractual and legal distance from operational use, and existing regulatory frameworks struggle to address such distributed responsibility, particularly where non-state actors play a central enabling role Chan (2018); Anstis (2024). This study is constrained by limited visibility into proprietary spyware systems and state intelligence workflows, does not examine specific downstream analytic models, and does not empirically evaluate outcomes. Future research could trace how data extracted through similar infrastructures are integrated into concrete analytic pipelines, including algorithmic prioritization or network inference systems, across institutional contexts.

6 Conclusion

This paper has argued that evaluations of AI dual-use risk in security and military contexts remain incomplete when confined to downstream algorithms or decision-support systems. Through an infrastructural analysis of commercial spyware and a case study of Pegasus, we have shown that upstream surveillance systems enable AI-mediated intelligence practices by automating persistent, covert data capture and stabilizing the data environments upon which analytic processes depend. Although Pegasus does not itself perform machine learning inference, it pre-conditions the feasibility and scale of downstream algorithmic analysis. Addressing AI-related security risks therefore requires extending accountability frameworks beyond models to include the surveillance architectures and privatized infrastructures that operationalize intelligence at scale.

ACKNOWLEDGEMENTS

We acknowledge the use of AI-based tools for language editing, and generating illustrative figures. All substantive claims, interpretations, and references remain the responsibility of the authors.

REFERENCES

Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, 20(3):973–989, 2018.

Siena Anstis. Regulating transnational dissident cyber espionage. *International & Comparative Law Quarterly*, 73(1):259–274, 2024.

Andrea Caruso. *Forensic Analysis of Mobile Spyware: Investigating Security, Vulnerabilities, and Detection Challenges in Android and iOS Platforms*. PhD thesis, Politecnico di Torino, 2024.

Anna W Chan. The need for a shared responsibility regime between state and non-state actors to prevent human rights violations caused by cyber-surveillance spyware. *Brook. J. Int'l L.*, 44:795, 2018.

Sidhant Chourasiya, Gyanesh Samanta, Devadarshan K Sardar, Ponnu Sharma, and CNS Vinoth Kumar. Pegasus spyware: A vulnerable behaviour-based attack system. In *2023 2nd international conference on edge computing and applications (ICECAA)*, pp. 287–292. IEEE, 2023.

Ron Deibert. Protecting society from surveillance spyware. *Issues in Science and Technology*, 38 (2):15–17, 2022.

Karwan Kareem. A comprehensive analysis of pegasus spyware and its implications for digital privacy and security. *arXiv preprint arXiv:2404.19677*, 2024.

Sean D Kaster and Prescott C Ensign. Privatized espionage: Nso group technologies and its pegasus spyware. *Thunderbird International Business Review*, 65(3):355–364, 2023.

Dan M Kotliar and Elinor Carmi. Keeping pegasus on the wing: legitimizing cyber espionage. *Information, Communication & Society*, 27(8):1499–1529, 2024.

Bill Marczak, John Scott-Railton, Noura Al-Jizawi, Siena Anstis, and Ron Deibert. The great ipwn: Journalists hacked with suspected nso group imessage 'zero-click'exploit. 2020.

Bill Marczak, John Scott-Railton, Bahr Abdul Razzak, Noura Al-Jizawi, Siena Anstis, Kristin Berdan, and Ron Deibert. Forcedentry: Nso group imessage zero-click exploit captured in the wild. 2021.

Donncha O'Cearbhaill and Bill Marczak. Exploit archaeology a forensic history of in the wild nso group exploits. In *Virus Bulletin Conference*, pp. 415, 2022.

JD Rudie, Zach Katz, Sam Kuhbander, and Suman Bhunia. Technical analysis of the nso group's pegasus spyware. In *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 747–752. IEEE, 2021.

Howie Shia. Forensic methodology report: How to catch nso group's pegasus. *Amnesty International*, 2021.

Brooke Spens. The spyware industrial complex. 2024.

Ana Pop Stefanija. Power asymmetries, epistemic imbalances and barriers to knowledge: the (im) possibility of knowing algorithms. In *Handbook of critical studies of artificial intelligence*, pp. 563–572. Edward Elgar Publishing, 2023.

Samuel Woodhams. *Spyware: An unregulated and escalating threat to independent media*. Center for International Media Assistance, 2021.