RISKS AND SAFETY CONSIDERATIONS FOR FOUNDATION MODEL-BASED AUTONOMOUS AGENTS' INTERACTION WITH THE ENVIRONMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation Model (FM) agents are increasingly deployed across diverse environments, from web automation to physical and medical systems. While their ability to interact autonomously enhances efficiency, it also introduces significant safety risks, including unauthorized access, data breaches, and system disruptions. Existing research on FM agent safety remains fragmented, lacking a comprehensive classification of risks across different domains. This paper addresses this gap by systematically categorizing risks into web, computer, and physical domains and proposing targeted mitigation strategies. Our framework aids researchers, developers, and policymakers in designing safer FM systems and establishing regulatory guidelines. By highlighting potential hazards and preventive measures, this work contributes to ensuring that FM agents operate securely while maximizing their transformative potential.

1 INTRODUCTION

Foundation Model (FM) agents, powered by advanced machine learning and natural language processing, are increasingly capable of interacting with digital and physical environments (Wang et al., 2024). These agents can perform a wide range of tasks, from automating web-based processes to controlling robotic systems and assisting in medical diagnostics (Moor et al., 2023; Firoozi et al., 2023). While such capabilities offer transformative potential, they also introduce significant safety concerns (Jabbour & Reddi, 2024). Agents that autonomously execute commands, access sensitive information, or make decisions in high-stakes environments introduce profound ethical and security dilemmas (de Cerqueira et al., 2024). But, *should an AI agent be allowed to bypass captchas, effectively sidestepping security mechanisms meant for human verification? If so, where do we draw the line between convenience and exploitation? Should these agents have access to financial credentials, potentially making autonomous transactions, or does that open the door to unprecedented fraud and abuse? Furthermore, how vulnerable are these systems to hacking or adversarial manipulation—could a malicious actor subtly nudge an agent toward harmful decisions without detection?* As FM agents become more integrated into critical infrastructure (McEvoy & Wolthusen, 2012), the risks of unintended consequences multiply. The complexity of their interaction with the environment demands a structured examination of these concerns, ensuring that autonomy does not come at the cost of security, privacy, or ethical integrity (Tang et al., 2024).

Despite growing interest in autonomous FM agents, research on their safety remains fragmented and domain-specific. Most studies either focus on ethical AI principles in general or address narrow technical vulnerabilities within specific applications (Osipov, 2024). However, a comprehensive framework classifying safety risks across multiple domains—such as web interactions, computer systems, physical systems, and the medical field—is lacking (Zeng et al., 2024). Without such a classification, it is challenging to develop effective safeguards, leading to vulnerabilities such as unauthorized access, data breaches, misconfigurations, and even physical harm (Shamsujjoha et al., 2024). This paper aims to fill this gap by systematically identifying the safety risks associated with FM agents' interaction with their environment and proposing targeted mitigation strategies. Two agenetic workflow examples and associated risks are added in Appendix A with Figure 3 and 4.

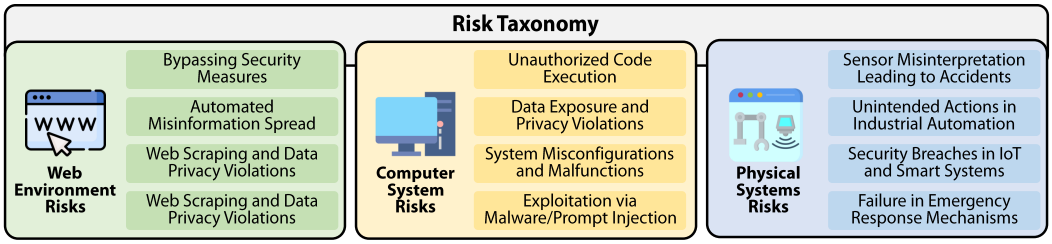

Figure 1: Our proposed **Risk Taxonomy** for FM Agents' Interaction with the Environment

Our contribution lies in defining a structured taxonomy of safety risks and offering a holistic perspective on potential hazards across different domains. By categorizing risks into web, computer, and physical environments, we provide a foundational framework that facilitates better risk assessment and policy development. This classification aids researchers, developers, and policymakers in designing safer FM systems, implementing necessary technical safeguards, and establishing regulatory guidelines. The implications of this work extend beyond theoretical insights, as it provides a practical roadmap for ensuring that FM agents enhance efficiency without compromising security, privacy, or human safety. We hope this framework fosters further research and development of robust safety mechanisms, enabling the responsible deployment of FM agents across various domains.

2 Risk Taxonomy

As FM agents gain the ability to interact with various environments, their autonomy introduces diverse safety concerns. These risks can be broadly categorized into four key domains, as illustrated in Figure 1: (1) web environment risks, where agents interact with online systems and may inadvertently bypass security mechanisms; (2) computer system risks, where executing commands or scripts could compromise data integrity; (3) physical systems risks, where misinterpretations or errors in autonomous control can lead to harm. This classification provides a structured approach to analyzing and mitigating potential hazards across different operational landscapes.

2.1 Web Environment Risks

Web environment risks arise when FM agents interact with online systems, potentially bypassing security measures, misusing web functionalities, or exposing sensitive data (Tóth et al., 2024). These agents may be programmed to perform automated tasks such as retrieving information, managing online accounts, or even navigating verification processes. However, their autonomy introduces security vulnerabilities that could be exploited or result in unintended consequences (Khan et al., 2024). For instance, an FM agent that automates captcha-solving could be repurposed to bypass website security features, allowing malicious actors to perform unauthorized transactions, spam content, or scrape sensitive data at scale (Bhowmick et al., 2023). Web environment risks can comprise:

Bypassing Security Measures: FM agents designed to fill in captchas or automate login processes may inadvertently violate website security policies. If exploited, this could lead to unauthorized access to restricted content, account takeovers, or data theft (Thakur et al., 2023).

Automated Misinformation Spread: Agents scraping web content to generate summaries or interact with social media could unintentionally amplify misinformation (Tomassi et al., 2024). If not properly filtered, such misinformation could harm reputations, sway public opinion, or cause economic losses.

Web Scraping and Data Privacy Violations: Some FM agents may extract personal or proprietary data from websites without consent, leading to legal and ethical concerns. For instance, an agent collecting financial data from online portals could expose sensitive user information (Schreyer et al., 2020).

Manipulation of Online Polls or Reviews: Autonomous agents capable of generating user interactions may be misused to manipulate product reviews, social media trends, or political polling results, thereby distorting genuine public opinion (Rosenberg, 2023).

2.2 Computer System Risks

These risks emerge when FM agents execute commands on local or remote computer systems, potentially affecting system integrity, accessing sensitive files, or disrupting normal operations. While such agents can automate software updates, run maintenance scripts, or optimize system performance, their unchecked execution can lead to security and privacy breaches (Cartrysse & van der Lubbe, 2003). For example, an FM agent tasked with automating system maintenance may accidentally execute a command that wipes a user's home directory, resulting in the loss of critical personal or professional files (Oks et al., 2006). Computer system risks can consist of:

Unauthorized Code Execution: If an FM agent has permission to execute terminal commands, it may run scripts that alter critical system files, disable security features, or corrupt important data, leading to system failure.

Data Exposure and Privacy Violations: An agent performing automated file management might access, copy, or share confidential data without proper authorization, resulting in personal or corporate data leaks.

System Misconfigurations and Malfunctions: Incorrect or incomplete execution of system commands could unintentionally disable essential services, delete necessary files, or cause software crashes, disrupting business operations.

Exploitation via Malware/Prompt Injection: If a malicious entity gains control over an FM agent using any type of malware or prompt injections, they could use it to download and install harmful software, including spyware, ransomware, or trojans, thereby compromising the entire system (Lee & Tiwari, 2024).

2.3 Physical Systems Risks

Physical system risks emerge when FM agents control real-world devices such as robots, autonomous vehicles, industrial machinery, or smart home systems. Errors in decision-making, misinterpretation of sensor data, or software failures could result in property damage, injury, or even loss of life. For instance, a warehouse robot controlled by an FM agent could misinterpret the position of a moving forklift, causing it to cross into its path and leading to a collision that damages both the robot and the warehouse equipment (Lehoux-Lebacque et al., 2024). Physical systems related risks may include

Sensor Misinterpretation Leading to Accidents: If an FM agent controlling an autonomous vehicle misinterprets environmental cues—such as stop signs, lane markings, or pedestrian movements—it could result in dangerous collisions (Fränzle & Hein, 2023).

Unintended Actions in Industrial Automation: Agents managing robotic arms or conveyor belts might miscalculate object positions, leading to product defects, equipment damage, or harm to workers operating nearby (Gouveia et al., 2024).

Security Breaches in IoT and Smart Systems: If an agent managing a smart home or factory control system is compromised, it could be remotely controlled by attackers, leading to security threats such as unauthorized door access or the disabling of critical safety mechanisms (Khanpara et al., 2023).

Failure in Emergency Response Mechanisms: FM agents deployed for security surveillance or hazard detection may fail to recognize emergency situations such as fire, gas leaks, or structural failures, leading to delayed responses and increased risk to human safety (Naim et al., 2021).

3 Safety Considerations and Mitigation Strategies

Ensuring the safe deployment of FM agents requires a comprehensive approach that balances automation with oversight, implements technical safeguards, and prioritizes data privacy and security, as outlined in Figure 2. As these agents increasingly interact with web systems, computer environments, physical devices, and medical applications, it is crucial to design frameworks that mitigate potential risks. This section outlines strategies to manage safety concerns while maximizing the benefits of FM agents.

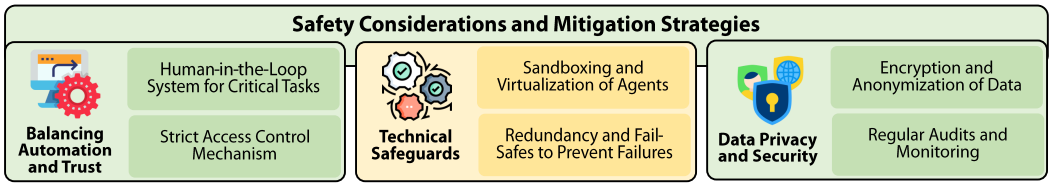

Figure 2: Safety Considerations and Mitigation Strategies

3.1 Balancing Automation and Trust

As FM agents become more autonomous, it is crucial to monitor their decision-making to prevent unintended consequences, as excessive autonomy without oversight can lead to security breaches, data loss, or physical harm. Balancing automation with human intervention ensures that critical tasks remain supervised, minimizing the risk of errors or malicious actions (Joseph et al., 2024). Integrating human supervision at key decision points, such as requiring human approval for system commands or medical diagnoses, is essential. For example, in healthcare, an FM agent should flag diagnoses for human review rather than making independent decisions, preventing irreversible mistakes (Ke et al., 2024). Additionally, implementing strict access controls ensures agents operate within defined limits, such as preventing agents from bypassing captchas or gaining unnecessary administrative privileges, reducing the risk of unauthorized actions or accidental system modifications.

3.2 Technical Safeguards

Technical safeguards are crucial to mitigating risks associated with FM agents, particularly when they interact with critical systems (Domkundwar et al., 2024). These measures help prevent unintended consequences by isolating agent operations, restricting their capabilities, and ensuring fail-safes are in place. For instance, sandboxing and virtualization keep FM agents in isolated environments, preventing them from directly accessing core system functions or sensitive data, which reduces the risk of system damage (Yedidia, 2024). In cybersecurity, this is useful for testing AI-driven automation without compromising system integrity. Additionally, redundancy and fail-safes are vital in physical systems. Using multiple sensors, setting operational limits, and adding emergency shutoff features ensure safety (Holst & Lohweg, 2021). For example, an FM agent controlling a robot should be equipped with collision detection, and autonomous vehicles should have manual override options to prevent accidents in case of errors.

3.3 Data Privacy and Security

Data privacy and security are crucial when FM agents handle sensitive information. Without proper safeguards, agents can expose personal data, violate regulations, or facilitate cyberattacks (Domkundwar et al., 2024). As these agents are deployed in sectors like finance, healthcare, and government, ensuring compliance with privacy laws is vital. Implementing encryption and anonymization ensures that data remains secure, with sensitive information being unreadable or removed before processing (Ajiga et al., 2024). For example, in healthcare, FM agents should analyze anonymized patient records to comply with regulations like HIPAA or GDPR (Ettaloui et al., 2023). Regular audits are also necessary to monitor agent activity, detect anomalies, and ensure security policies are followed, with automated systems flagging suspicious behaviors or unauthorized access.

4 Conclusion

In conclusion, as Foundation Model agents become more integrated into various domains, ensuring their safe operation is crucial to prevent unintended consequences. By implementing a combination of human oversight, technical safeguards, and strict data privacy protocols, we can mitigate the risks associated with their interaction with web, computer, physical, and medical systems. These strategies help strike a balance between automation and security, ensuring FM agents enhance productivity and efficiency without compromising safety or ethical standards. Ultimately, these efforts will pave the way for responsible, secure, and effective deployment of autonomous systems.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2025 Tiny Papers.

References

Daniel Ajiga, Patrick Azuka Okeleke, Samuel Olaoluwa Folorunsho, and Chinedu Ezeigweneme. Designing cybersecurity measures for enterprise software applications to protect data integrity, 2024.

Rajat Subhra Bhowmick, Rahul Indra, Isha Ganguli, Jayanta Paul, and Jaya Sil. Breaking captcha system with minimal exertion through deep learning: Real-time risk assessment on indian government websites. *Digital Threats: Research and Practice*, 4(2):1–24, 2023.

K Cartrysse and JCA van der Lubbe. Providing privacy to agents in an untrustworthy environment. *system*, 38:39, 2003.

José Antonio Siqueira de Cerqueira, Mamia Agbese, Rebekah Rousi, Nannan Xi, Juho Hamari, and Pekka Abrahamsson. Can we trust ai agents? an experimental study towards trustworthy llm-based multi-agent systems for ai ethics. *arXiv preprint arXiv:2411.08881*, 2024.

Ishaan Domkundwar, Ishaan Bhola, et al. Safeguarding ai agents: Developing and analyzing safety architectures. *arXiv preprint arXiv:2409.03793*, 2024.

Nehal Ettaloui, Sara Arezki, and Taoufiq Gadi. An overview of blockchain-based electronic health record and compliance with gdpr and hipaa. In *The International Conference on Artificial Intelligence and Smart Environment*, pp. 405–412. Springer, 2023.

Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, pp. 02783649241281508, 2023.

Martin Fränzle and Andreas Hein. Safer than perception: Increasing resilience of automated vehicles against misperception. In *International Conference on Bridging the Gap between AI and Reality*, pp. 415–433. Springer, 2023.

Eber L Gouveia, John G Lyons, and Declan M Devine. Implementing a vision-based ros package for reliable part localization and displacement from conveyor belts. *Journal of Manufacturing and Materials Processing*, 8(5):218, 2024.

Christoph-Alexander Holst and Volker Lohweg. A redundancy metric set within possibility theory for multi-sensor systems. *Sensors*, 21(7):2508, 2021.

Jason Jabbour and Vijay Janapa Reddi. Generative ai agents in autonomous machines: A safety perspective. *arXiv preprint arXiv:2410.15489*, 2024.

Sunday Joseph, Titilayo Modupe Kolade, Onyinye Obioha Val, Olubukola Omolara Adebiyi, Olumide Samuel Ogungbemi, and Oluwaseun Oladeji Olaniyi. Ai-powered information governance: Balancing automation and human oversight for optimal organization productivity. *Asian Journal of Research in Computer Science*, 17(10):10–9734, 2024.

Yu He Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. Enhancing diagnostic accuracy through multi-agent conversations: Using large language models to mitigate cognitive bias. *arXiv preprint arXiv:2401.14589*, 2024.

Raihan Khan, Sayak Sarkar, Sainik Kumar Mahata, and Edwin Jose. Security threats in agentic ai system. *arXiv preprint arXiv:2410.14728*, 2024.

Pimal Khanpara, Kruti Lavingia, Rajvi Trivedi, Sudeep Tanwar, Amit Verma, and Ravi Sharma. A context-aware internet of things-driven security scheme for smart homes. *Security and Privacy*, 6 (1):e269, 2023.

Donghyun Lee and Mo Tiwari. Prompt infection: Llm-to-llm prompt injection within multi-agent systems. *arXiv preprint arXiv:2410.07283*, 2024.

Vassilissa Lehoux-Lebacque, Tomi Silander, Christelle Loiodice, Seungjoon Lee, Albert Wang, and Sofia Michel. Multi-agent path finding with real robot dynamics and interdependent tasks for automated warehouses. In *ECAI 2024*, pp. 4393–4401. IOS Press, 2024.

Thomas McEvoy and Stephen Wolthusen. Agent interaction and state determination in scada systems. In *Critical Infrastructure Protection VI: 6th IFIP WG 11.10 International Conference, ICCIP 2012, Washington, DC, USA, March 19-21, 2012, Revised Selected Papers 6*, pp. 99–109. Springer, 2012.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Aryan Naim, Ryan Alimo, and Jay Braun. Ai agents in emergency response applications. *arXiv preprint arXiv:2109.04646*, 2021.

Artem A Oks, Hanumantha Rao Kodavalla, and Martin J Sleeman. Systems and methods for automated maintenance and repair of database and file systems, November 28 2006. US Patent 7,143,120.

Daniil V Osipov. Ethics of ai technologies in "sensitive" content creation and evaluation. school shooting cases. *Galactica Media: Journal of Media Studies*, 6(3):44–65, 2024.

Louis Rosenberg. The manipulation problem: conversational ai as a threat to epistemic agency. *arXiv preprint arXiv:2306.11748*, 2023.

Marco Schreyer, Chistian Schulze, and Damian Borth. Leaking sensitive financial accounting data in plain sight using deep autoencoder neural networks. *arXiv preprint arXiv:2012.07110*, 2020.

Md Shamsujjoha, Qinghua Lu, Dehai Zhao, and Liming Zhu. Towards ai-safety-by-design: A taxonomy of runtime guardrails in foundation model based systems. *arXiv e-prints*, pp. arXiv–2408, 2024.

Xiangru Tang, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, Yilun Zhao, Jian Tang, Zhuosheng Zhang, et al. Prioritizing safeguarding over autonomy: Risks of llm agents for science. *arXiv preprint arXiv:2402.04247*, 2024.

Ashutosh Thakur, Bhavishya, and Priya Singh. A novel deep learning based fully automated framework for captcha security vulnerability checking. In *International Conference on Data Science and Network Engineering*, pp. 431–443. Springer, 2023.

Andrea Tomassi, Andrea Falegnami, and Elpidio Romano. Mapping automatic social media information disorder. the role of bots and ai in spreading misleading information in society. *Plos one*, 19(5):e0303183, 2024.

Rebeka Tóth, Tamas Bisztray, and László Erdődi. Llms in web development: Evaluating llm-generated php code unveiling vulnerabilities and limitations. In *International Conference on Computer Safety, Reliability, and Security*, pp. 425–437. Springer, 2024.

Shuai Wang, Weiwen Liu, Jingxuan Chen, Weinan Gan, Xingshan Zeng, Shuai Yu, Xinlong Hao, Kun Shao, Yasheng Wang, and Ruiming Tang. Gui agents with foundation models: A comprehensive survey. *arXiv preprint arXiv:2411.04890*, 2024.

Zachary Yedidia. Lightweight fault isolation: Practical, efficient, and secure software sandboxing. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pp. 649–665, 2024.

Yi Zeng, Kevin Klyman, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. Ai risk categorization decoded (air 2024): From government regulations to corporate policies. *arXiv preprint arXiv:2406.17864*, 2024.

A Foundation Model (FM) Agents in the Wild: Risk Examples

The examples in Figure 3 and 4 highlight the broad spectrum of risks posed by FM agents across different domains. The financial analyst agent demonstrates how automation, if unchecked, can overstep permissions, execute unintended trades, and expose users to financial losses or compliance violations. Meanwhile, the media agent exposes severe privacy risks, where aggregating publicly available but scattered data can reconstruct an individual's personal life, leading to doxxing, harassment, and misinformation spread. These examples show how autonomous agents can unintentionally breach security, ethical, and regulatory boundaries when interacting with online systems. The risks become even more concerning when extended to everyday users, as similar pipelines could be used to scrape personal information from social media without consent, raising serious concerns about digital privacy, data ethics, and AI governance.

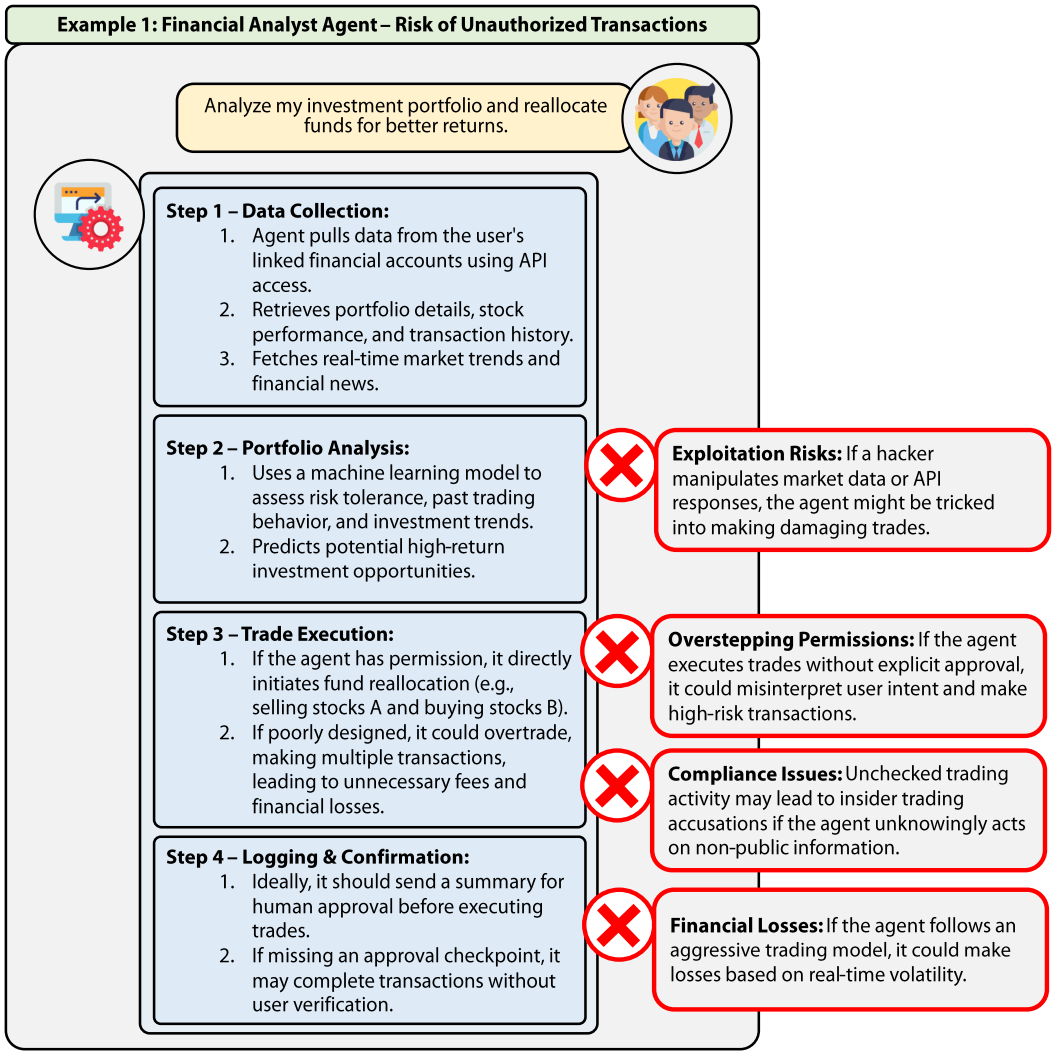

Figure 3: Example 1: Financial Analyst Agent – Risk of Unauthorized Transactions (Computer and Physical system Risks)

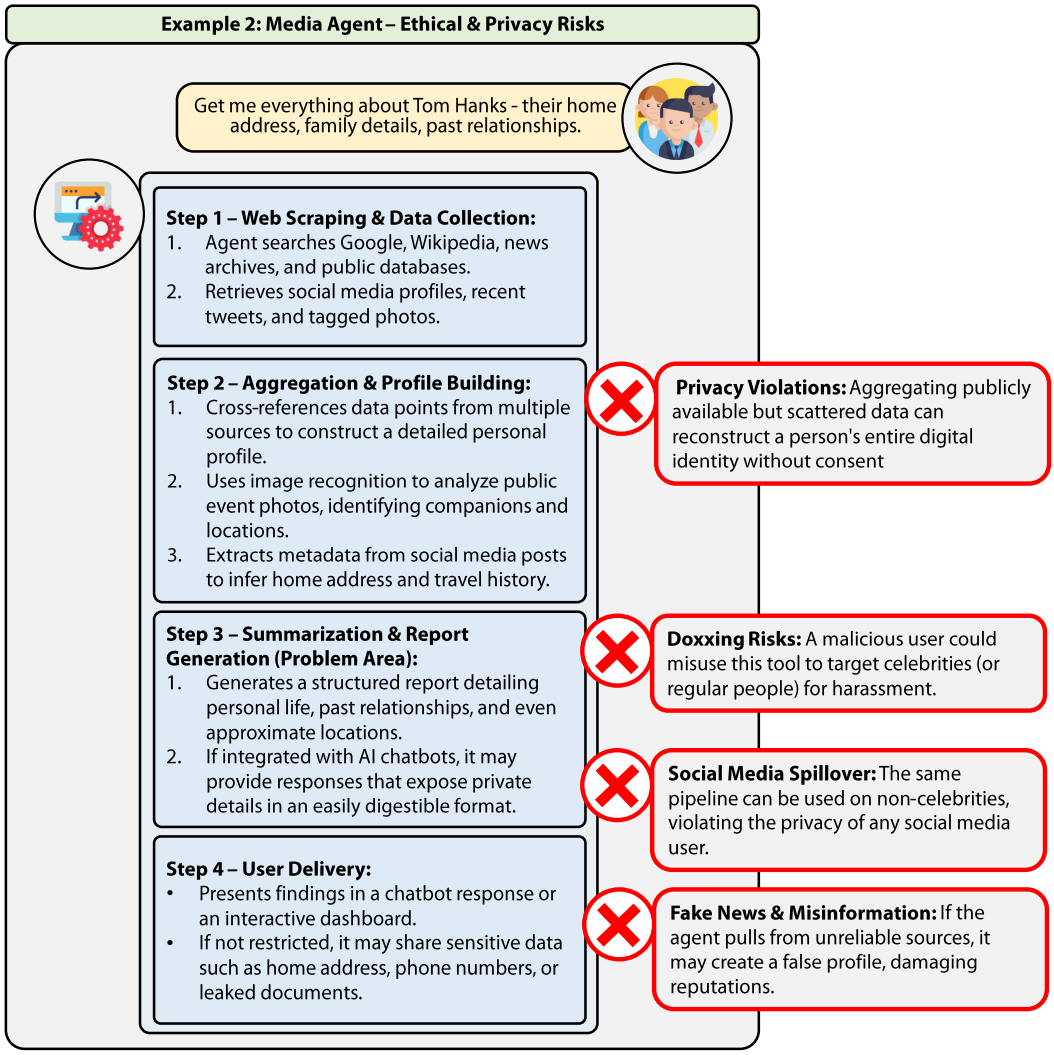

Figure 4: Example 2: Media Agent – Ethical and Privacy Risks (Web Environment Risk)

These examples highlight the broad spectrum of risks posed by FM agents, aligning with our Risk Taxonomy. The financial analyst agent falls under computer system risks, as it demonstrates how automation, if unchecked, can overstep permissions, execute unintended trades, and expose users to financial losses or compliance violations. By gaining direct access to financial tools and transaction systems, such an agent risks unauthorized code execution, data exposure, and system misconfigurations, which could have significant economic and security repercussions. Meanwhile, the media agent aligns with web environment risks, as it exposes severe privacy violations. By aggregating publicly available but scattered data, the agent reconstructs an individual's personal life, leading to doxxing, harassment, and misinformation spread. This falls under web scraping and data privacy violations, as well as automated misinformation spread, showing how AI-driven agents can unintentionally breach ethical, security, and regulatory boundaries. Extending these concerns to everyday social media users, similar pipelines could scrape personal data—including images, relationships, and location details—without consent, escalating concerns about digital privacy, data ethics, and AI governance.