Securing People by Hybrid Offensive Security: A Preliminary Case Study for Social Safety

Vita Santa Barletta[1,†], Danilo Caivano[1,2,†], Samuele del Vescovo[3,*,†], Giuseppe Fenza[4,†], Vincenzo Loia[4,†] and Pio Pasquale Trotta[3,4,*,†]

[1]*Software Engineering Research LABoratory (SERLAB), Via Orabona 4, 70125 Bari, Apulia, Italy*
[2]*Consorzio Interuniversitario Nazionale per l'Informatica (CINI)*
[3]*IMT School for Advanced Studies Lucca, Piazza S.Francesco, 19, 55100 Lucca, Italy*
[4]*Department of Management and Innovation Systems, University of Salerno, Fisciano (SA), 84084, Italy*

Abstract

This research work aims to advance the understanding of how disinformation-driven social attacks and cyber attacks targeting critical AI-enabled systems can converge within complex hybrid threat landscapes. The broader objective is to examine the interdependencies between attacks operating in the *informational* domain and intrusions affecting cyber–physical infrastructures, with a specific focus on the connected and automated vehicles, under the *Cyber* attacks set. In this scenario, the *informational* attack exploits disinformation concerning the malfunctioning or unreliability of vehicular components, thereby fostering public distrust and shaping cognitive responses at both individual and societal levels. In a semi-concurrent fashion, the *Cyber* attack can target a Machine Learning (ML) based Intrusion Detection System (IDS) useful to detect Controller Area Network (CAN) malicious frames. By employing an adversarial machine learning approach, the attack seeks to compromise the IDS and, in doing so, degrade or disable critical safety functions of the vehicle by a vulnerable electronic control unit. To model the previous convergence, the proposed work introduces a possible hybrid threat model, grounded on a hybrid kill chain. Ultimately, assessing the potential negative impacts from the broader perspective of hybrid risk management can contribute to the development of a cyber social defense posture.

Keywords

Disinformation, Information Disorder Awareness, Automotive Cybersecurity, Adversarial Machine Learning, Cyber Social Security, Hybrid Threat

CEUR-WS.org/Vol-4329/paper14.pdf

1. Introduction

1.1. Motivation

The current international geopolitical scenario is marked by profound instability, creating an unprecedented climate of uncertainty in which the escalation of intercontinental tensions into broader conflict appears increasingly plausible [1]. This volatility fosters a permissive setting for the proliferation and interaction of both conventional threats and hybrid challenges [2, 3], enabled by rapid technological change and intensifying competition across energy, cyber, and information domains.

Hybrid threats deliberately combine military instruments with non-kinetic tactics that exploit political, societal, and technological vulnerabilities [4, 2]. A relevant conceptual frame in hybrid scenarios is the multi-domain operations paradigm [2], where synchronized capabilities across the *Cyber*, *Land*, *Space*, *Sea*, and *Air* domains generate synergistic effects that overwhelm a victim's capacity for comprehension. Additional *informational* and *electromagnetic* environments further enable these capabilities.

1st Workshop on Symbiotic AI for Cyber Social Security: Human–AI Collaboration for Social Resilience and Trustworthy Digital Ecosystems (SAI4CCS 2026), co-located with the 2nd Joint National Conference on Cybersecurity (ITASEC & SERICS 2026), February 09–13, 2026, Cagliari, Italy

*Corresponding author.
†These authors contributed equally.
vita.barletta@uniba.it (V. S. Barletta); danilo.caivano@uniba.it (D. Caivano); samuele.delvescovo@imtlucca.it (S. d. Vescovo); gfenza@unisa.it (G. Fenza); loia@unisa.it (V. Loia); piopasquale.trotta@imtlucca.it (P. P. Trotta)
0000-0002-0163-6786 (V. S. Barletta); 0000-0001-5719-7447 (D. Caivano); 0009-0001-8220-5135 (S. d. Vescovo); 0000-0002-4736-0113 (G. Fenza); 0000-0003-4807-8942 (V. Loia); 0009-0001-0735-3939 (P. P. Trotta)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-14

Within this context, disinformation campaigns play a central role [5]. Their aim is to manipulate perceptions, erode institutional trust, and distort decision-making. Disinformation constitutes a deliberate and structured process involving the fabrication, orchestration, and amplification of misleading content to induce cognitive overload [5, 2], thereby complicating the ability of individuals and institutions to distinguish trustworthy from unreliable information [6, 7]. Civilian populations often represent the most vulnerable targets [7].

Emerging smart cities exemplify this vulnerability [8, 9]. These environments heavily rely on Connected and Automated Vehicles (CAVs) integrated into broader Connected and Automated Mobility (CAM) systems [10], which constitute potential critical infrastructures. Despite their sophistication, CAV ecosystems exhibit systemic cyber vulnerabilities [11], as their dependence on continuous connectivity and extensive data exchange expands the attack surface [12, 13]. Some of these are based on the Controlled Area Network (CAN) protocol [14]. From a cyber perspective, Machine Learning (ML)-based Intrusion Detection Systems (IDSs) are a possible countermeasure for the in-vehicle networks security [15, 16]. However, these represent particularly exposed targets, being susceptible to adversarial manipulations such as Adversarial Machine Learning (AML) attacks [17, 18, 19].

In such settings, disinformation and cyber attacks can be jointly orchestrated. A coordinated scenario may involve disinformation campaigns fabricating accident reports or warnings of imminent failures, while simultaneous cyber attacks exploit CAV vulnerabilities to cause real disruptions. This convergence would blur the distinction between genuine failures and fabricated narratives, undermining public trust and complicating crisis-response efforts.

Understanding these hybrid dynamics is essential for developing a cyber social threat intelligence, namely an expanded conception of threat intelligence that moves beyond a purely *Cyber* focus. While the disinformation lifecycle [5] offers a foundational framework, further empirical research is required to examine how societies respond when false narratives coincide with physical disruptions. Key open questions concern whether populations become more susceptible to manipulative messaging during technological failures, or conversely, whether visible disruptions increase skepticism toward disinformation. Addressing such questions is crucial for designing robust forms of cognitive resilience that extend beyond media literacy to encompass crisis behavior in digitally mediated societies.

1.2. Research Goal

Considering the previous motivations, the Research Goals (RGs) of this work can be summarized as follows:

1. RG1: Explore a possible hybrid offensive security strategy that combines disinformation with a cyber operation aimed at compromising the reliability of a ML based IDS for CAN bus detection;
2. RG2: Design an example of an hybrid threat model integrating disinformation and automotive cyber attack exploits;
3. RG3: Design an example of an hybrid kill chain showing the symbiotic and resonant dynamics between these attack types.

1.3. Research Question

The following Research Questions (RQs) are extracted from the RGs:

- RQ1: How can an *informational* threat be effectively combined with a *Cyber* automotive threat in a mutually resonant way?
- RQ2: What could be an example of hybrid threat model for the aforementioned hybrid threat?
- RQ3: How can a kill chain for *informational* threats be effectively combined with *Cyber kill chains* (in the automotive sector)?

1.4. Contribution

The contribution of this work are:

1. Examine an illustrative hybrid offensive strategy that combines pre-emptive disinformation with a semi-parallel cyber operation designed to degrade the reliability of a CAN-bus ML-based IDS;
2. Explore plausible an illustrative hybrid threat model that integrates disinformation operations with automotive-focused cyber attacks;
3. Explore plausible tactics contributing to an hybrid kill chain that demonstrates the symbiotic dynamics between these informational and cyber-technical attack modalities.

2. Related Work

2.1. Hybrid Conflicts: The New Paradigm

Hybrid threats represent a particularly complex scenario of contemporary conflict [20, 21]. They integrate diverse instruments, ranging from cyber attacks and disinformation to sabotage, economic coercion, and kinetic actions, into coordinated operations that are difficult to manage, attribute, and counter [2]. Their strategic effectiveness derives from deliberate ambiguity, enabling adversaries to operate below conventional thresholds of conflict and exploit systemic vulnerabilities [22].

A central component of these operations is the manipulation of the *informational* environment [2]. Disinformation functions not as an auxiliary tactic but as a core capability, frequently deployed alongside *Cyber* or physical actions. Recent conflicts illustrate this dynamic: during the 2025 India–Pakistan crisis, fabricated reports of downed aircraft, political instability, and attacks on civilian areas circulated widely on social and mainstream media, generating confusion and inflaming public sentiment [23]. Similar patterns have emerged in Europe, where drone incursions and the use of "shadow fleets" have been accompanied by targeted disinformation campaigns intended to erode trust and social cohesion [24].

Hybrid operations also blend technical disruption with psychological pressure, as seen during the COVID-19 pandemic, when cyber attacks on already-strained healthcare infrastructures amplified societal anxieties [25, 8]. These examples illustrate the growing prominence and sophistication of hybrid tactics across contemporary security environments.

Hybrid Offensive Security In line with RQ1, the literature consistently identifies three recurrent strategic vectors: (1) cyber-physical attacks on critical infrastructure, (2) integrated *informational* and *Cyber* operations, and (3) supply-chain and software-based compromises. The 2015 Ukrainian power-grid incidents exemplify the first approach: staged phishing campaigns, lateral movement into control networks, destructive malware deployment, and manipulation of SCADA systems resulted in prolonged outages and impaired emergency response [26, 26]. Empirical studies of state-linked campaigns, such as Russian operations targeting the United States [27] and Ukraine [28], demonstrate the powerful union between narrative engineering and technical offenses. Canonically, disinformation acts as a force multiplier by complicating attribution, amplifying the effects of technical disruptions, and eroding societal trust. A critical vector for large-scale and deniable operations are the supply-chain compromission [29]. Canonical cases such as SolarWinds and NotPetya illustrate how malicious updates or trojanized components can propagate broadly through trusted software ecosystems. The operational toolkit documented in the literature spans phishing, credential theft, wipers and ransomware, trojanized updates, compromised build pipelines [29]. However, defining hybrid threat models remains challenging due to the diffuse and evolving nature of their attack surfaces. Consequently, this work focuses on a specific illustrative case considering the automotive sector.

Hybrid Defensive Security Contemporary literature frames defensive approaches to hybrid conflict as inherently cross-domain and multidisciplinary ones, requiring: (1) technical resilience and detection, (2) *informational* countermeasures, and (3) organisational and legal preparedness. The first approach

includes the fortification of IT and OT infrastructures, robust incident-response capabilities, and the deployment of defences against adversarial manipulation [30, 31]. Regarding *informational* countermeasures, frameworks such as the Council of Europe's Information Disorder model highlight layered approaches i.e., prebunking, timely fact-checking, and amplification of credible sources, as essential to slowing the spread of harmful content [32]. Operationalizing hybrid-threat intelligence remains profoundly challenging. Data interoperability and cross-sector information sharing are constrained by legal, commercial, and privacy barriers, and standardized taxonomies for disinformation incidents remain underdeveloped [33]. Simultaneously, adversaries continue to evolve their TTPs, leveraging criminal intermediaries and AI-generated content to accelerate and scale hybrid campaigns [34].

2.2. Information Disorder in Hybrid Scenario

The contemporary security environment is defined by hybrid conflict, a strategy that seamlessly integrates traditional military force with pervasive information disorder and cyber tactics. Following the Council of Europe's framework, this disorder comprises misinformation, disinformation, and malinformation, collectively constituting "information pollution" amplified by digital platforms [35]. These operations are strategic instruments, often state-orchestrated, designed to achieve geopolitical objectives by targeting civilian and military cognitive domains, thereby eroding trust, manipulating public opinion, and undermining decision-making, a process the EU terms Foreign Information Manipulation and Interference (FIMI) [36]. Disinformation campaigns are thus a powerful asymmetric weapon, blurring into psychological operations and leveraging cyber tools for rapid, deceptive spread [37]. Case studies, most recently the Russia–Ukraine conflict, reveal the systematic synchronization of massive cyber attacks with intense, state-run propaganda to demoralize societies and justify aggression [38]. Globally, state actors are projecting influence by converging narratives (e.g., China–Russia cooperation) and exploiting local fault lines with deepfakes and AI-generated content to polarize democracies. Ultimately, these hybrid threats profoundly impact civilian security, making the society resilience, media literacy, and fact-checking efforts critical defense mechanisms.

2.3. Smart City & Automated Vehicles in Hybrid Scenario

Smart cities leverage digital technologies to improve efficiency, sustainability, and overall quality of life [39]. These ecosystems integrate critical infrastructures for transportation, public services, healthcare, and smart homes, all interconnected through IoT-based architectures [40]. Within this broader system, Connected and Automated Vehicles (CAVs) operate as networked platforms equipped with advanced communication and ML capabilities that support infrastructure interaction, and inter-vehicle coordination [10, 41]. CAVs remain vulnerable to multiple *Cyber* attack surfaces, particularly those associated with the CAN network [14], including Denial of Service (DoS), Fuzzing, Replay, and Spoofing attacks [11]. Intrusion Detection Systems (IDSs) are deployed to identify such anomalies by monitoring deviations in CAN traffic features. ML-based IDSs, which analyze attributes such as the CAN Identifier (ID), payload content, and physical-layer characteristics, span both unsupervised and supervised approaches [42, 43, 44].

However, these systems are susceptible to Adversarial Machine Learning (AML) attacks, particularly evasion attacks that generate adversarial examples, namely inputs that remain nearly indistinguishable from benign samples yet induce misclassification during validation or deployment. While white-box attacks dominate the existing literature [45, 46, 47], black-box settings are especially concerning, as adversaries require no knowledge of the target model's internal structure.

To the best of our knowledge, the literature lacks comprehensive analyses of purely adversarial and polymorphic black-box evasion attacks targeting ML-based IDSs for CAN buses. Although white-box studies suggest that extending these methods to black-box scenarios is feasible and prior demonstrations of attacks on unprotected CAN networks support the specific risks such techniques pose within hybrid conflict contexts remain largely unexplored. Moreover, to the best of our knowledge, no existing studies examine methodologies for coupling *informational* operations with cyber attacks exploiting

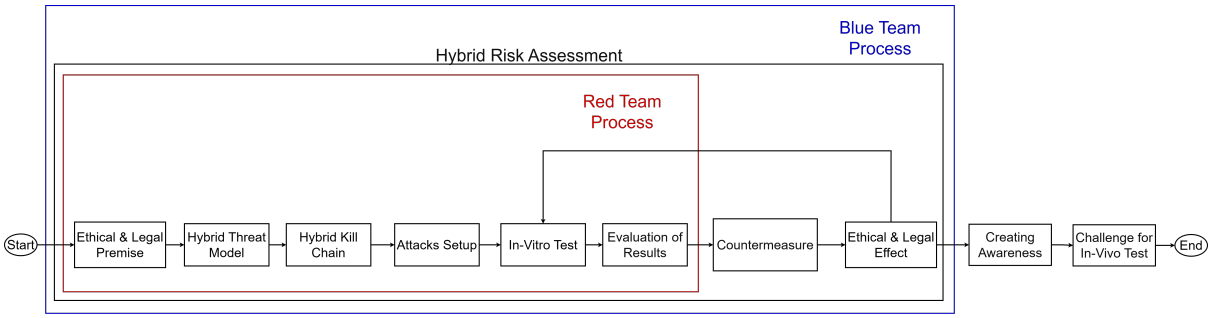


Figure 1: Methodology

vulnerabilities in CAN-based in-vehicle networks within the automotive domain. Even though in [48], a similar campaign targeting a victim bank is described, this gap is particularly consequential given the profound ongoing transformations in the sector.

3. Methodology

To address the RQ1 and fulfill the RG1, it is important to consider the timing of these sub-operations. Time can be the principal dimension for evaluating how the two attack "modalities" may be combined: the *Cyber* operation may precede, coincide with, or follow the disinformation campaign. If the *Cyber* intrusion precedes the *informational* campaign, the latter functions primarily as an amplifier without requiring precise temporal synchronization; if it follows, the combined rationale is limited. Therefore, this study focuses on the scenario in which the *Cyber* and *informational* vectors are semi-concurrent (i.e. very close in time), examining their interdependencies. The analysis mainly considers the scenario in which the *Cyber* component is exploited to support the disinformation campaign, amplifying its effects. All this is done while acknowledging certain methodological limitations. Figure 1 shows the methodology. Assessing the impact of this offense plays an important role in this methodology, as it allows us to understand how disastrous it could be. The red team (framed into a more complex architecture of blue team) plays a central role in threat analysis and in the development of effective countermeasures. The initial phase, examining the ethical and legal premises, is critical. It delineates the normative boundaries and operational constraints within which red-team activities must occur. The hybrid threat model and hybrid kill chain are useful in order to understand the capabilities and step performable by attackers. The outcomes of in-vivo testing provide valuable insights into the experimental feasibility of the envisaged actions. These findings subsequently inform threat-intelligence processes by identifying corresponding defensive measures which, once implemented, can be leveraged to mitigate the risks posed by such threats. Part of these are the ethical and legal effects. Finally, responsible disclosure remains essential for raising awareness and fostering a robust cyber-social security posture.

4. Hybrid Offense Orchestration

4.1. Hybrid Threat Model

Figure 2 shows the hypothesized hybrid threat model, answering to RQ2 and and fulfilling the RG2. Over the temporal dimension, the hybrid threat model posits that an hybrid adversary (the “hybrid red team” represented by the cloud) may execute the *informational* and *Cyber* sub-offensives in a semi-concurrent fashion. Such a team is envisaged as multidisciplinary: cyber security specialists are responsible for penetrating and subverting technical attack surfaces associated with connected and autonomous vehicles (e.g., CAN and IDS network), while psychologists and communication experts design and steer the disinformation campaign to maximise cognitive impact on civilian audiences and

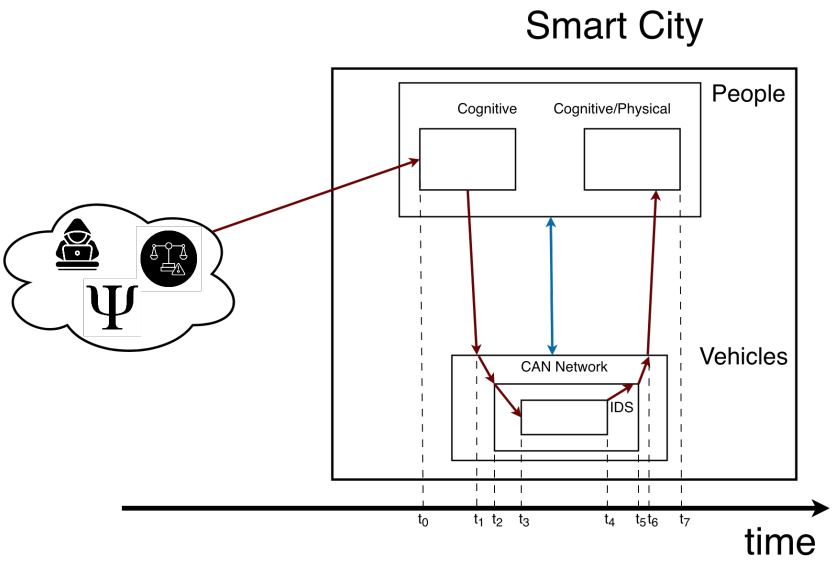


Figure 2: Hybrid threat model hypothesized

to amplify the consequences of the cyber intrusion. Legal specialists play a critical strategic role by identifying and exploiting regulatory grey zones at national and international levels, thereby extending the operational scope and persistence of the offensive. The red arrows denote the attack vectors. The primary vector is *informational*, directed at generating cognitive disarray among the civilian population by alleging or amplifying perceived malfunctions in particular vehicles. A few moments later (or at the same time) , the hybrid red team advances a *Cyber* sub-offensive that exploits technical vulnerabilities. Successful *Cyber* intrusion can reinforce and magnify the informational campaign, producing reciprocal amplification among the various actors within this constructed socio-technical scenario. The blue arrow represents the plausible interaction between citizens and connected/autonomous vehicles, both of which function as sub-assets within the smart-city environment. Rectangular nodes identify the distinct attack surfaces associated with each asset. It is possible to distinguish an *informational* offensive that precedes the *Cyber* intrusion, followed by its continuation in the post-intrusion phase, thereby making explicit the symbiotic relationship between the two sub-offensives.

4.2. Hybrid Kill Chain

To answer to RQ3 and fulfill the RG3, the hypothesized hybrid kill chain conceptually resembles a Russian nesting doll. Figure 3 shows the hybrid kill chain. The attack hypothesized and described in this work presents three attack "flows" over time. The first one is represented by the *informational* attack. During the "execute" phase of the relative kill chain, the *Cyber* attack applied to a family of vehicles can begin; also, the *informational* and *cyber* offensives may be semi-concurrent. During this last attack stream, starting from the "installation" phase (at time two), the attack on the IDS can be contextualized. At the end of this, it is possible to trace the "stack" of attacks representing the hybrid attack flow. The symbiotic relationship between the sub–kill chains is manifested in the attack flow, where they coexist in a mutually reinforcing manner while retaining distinct functions within a shared temporal progression. This final aspect elucidates the resonant effect of these tactics: beginning from the most deeply embedded "layer," their upward progression through successive tactical phases generates cumulative hybrid effects, thereby magnifying the overall impact.

4.2.1. Informational Kill Chain

To construct a detailed *informational* kill chain for this component of the hybrid offensive, adversary movements are mapped to Tactics, Techniques, and Procedures (TTPs) using the DISARM framework

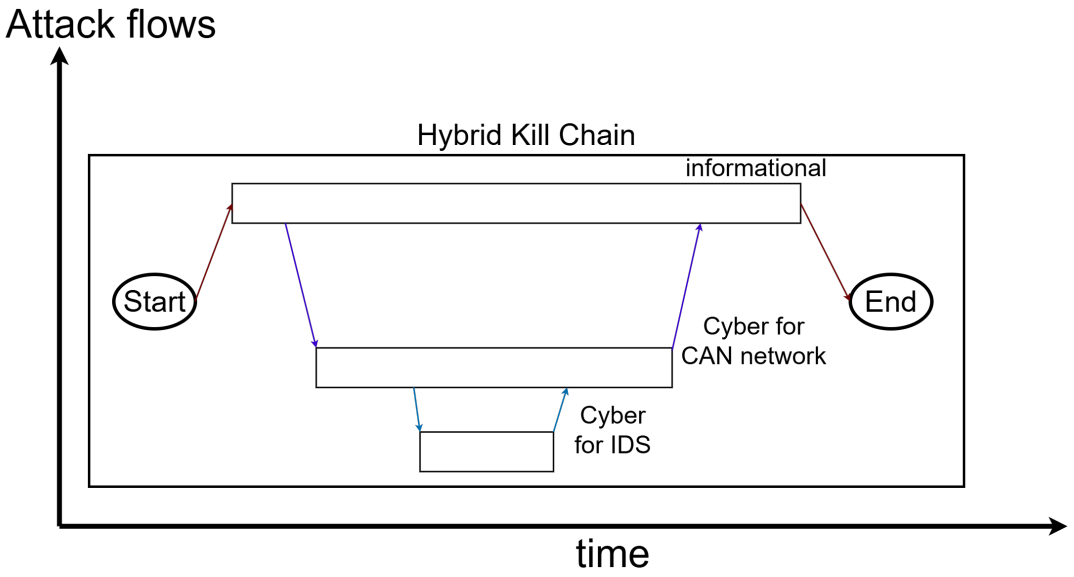


Figure 3: Hybrid kill chain hypothesized.

[49].

During the "plan" phase, the attacking organization establishes the overarching objectives and strategic framework of the campaign preceding the *Cyber* attack. Specifically, it defines the intended outcomes of the disinformation campaign and the set of conditions necessary to achieve these objectives.

The ultimate aim is to foster social disorder, through the amplification of ideological divisions among subgroups within targeted communities, and to induce psychological distress, thereby compromising the reputational integrity of automotive manufacturers and threatening national security [50]. Ultimately, the pursuit of dominance and sovereignty by states or organizations serves as the principal driving force behind such actions.

In the "prepare" phase, the criminal organization constructs new narratives designed to advance the strategic objectives of the disinformation campaign, aligning them with the broader goals of the intended operations [50].

Nonetheless, the creation of original narratives, particularly those of a conspiratorial nature, typically demands considerably greater effort than the adaptation or repurposing of pre-existing ones. Notably, the heightened attention of traditional and social media to a particular news item, when coupled with the limited knowledge or incomplete understanding of the targeted audience, serves to amplify speculation, rumor propagation, and the spread of conspiracy theories—all of which are highly susceptible to manipulation by malicious actors [50].

An essential component of these narratives consists of posts, captions, and comments that exploit the indexing mechanisms of social media correlation algorithms, with hashtags playing a crucial amplifying role. Dissemination through major social networking platforms, strategically intensified via targeted hashtags, seeks to expand the reach of the offensive by leveraging clickbait tactics to capture user attention, thereby fostering polarization within affected communities and contributing to the formation of potential echo chambers.

During the "execute" phase, the criminal organization seeks to disseminate its constructed narratives, endeavoring to enhance their perceived credibility by leveraging pre-existing narratives that align with those previously established to support the overarching objectives of the operation.

In the "assessment" phase, the objective is to employ one or more metrics (or indicators) to evaluate the current condition of the affected social system. In this context, the criminal organization aims to identify and continuously monitor these indicators throughout the progression of the disinformation campaign, thereby enabling the measurement of its performance and effectiveness both during its execution and in the aftermath of its real-world application. Figure 4 shows the aforementioned kill

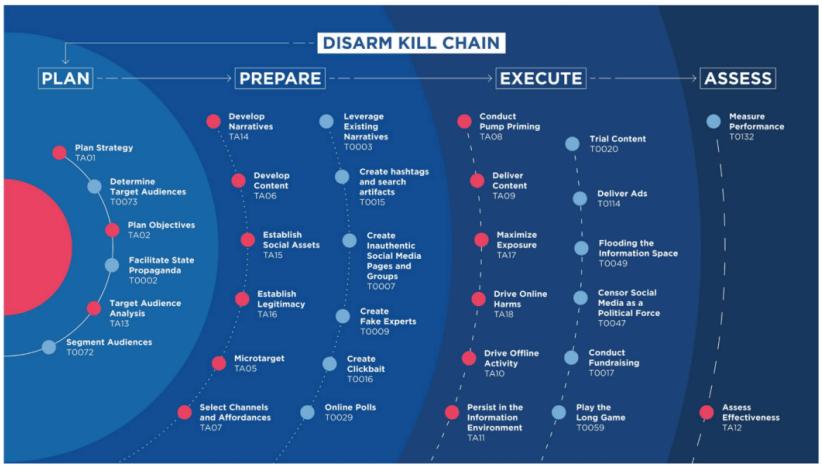


Figure 4: *informational* kill chain based in DISARM Framework [49]

chain.

As proactive coutermeasure, ML-based disinformation detectors leveraging advanced natural language processing and network analysis to identify deceptive patterns at scale can be adopted [51]. By integrating the Framework to Create Symbiotic Artificial Intelligence (Frame SAI) [52], these systems transition from mere automation to human-AI symbiosis, fostering a bidirectional relationship that supports rather than replaces human experts. These detectors adhere to Guideline 1.1 by enabling human-in-the-loop oversight, ensuring operators can check, monitor, and supervise the system's behavior. Furthermore, by implementing Success Criteria 1.2.3, such systems provide interfaces for humans to revise AI decisions through thresholds or rules, ensuring that final judgment remains a human-led, transparent, and legally compliant process [53].

4.2.2. Cyber Kill Chain

To construct a detailed *Cyber* kill chain for this component of the hybrid offensive, adversary movements are mapped to TTPs using the MITRE ATT&CK framework [54]. This mapping is then augmented by the Advanced Threat Knowledge-base for Networked Automotive (ATHENA) model [55] and by MITRE ATLAS [56] to represent automotive and ML-specific attack primitives and AML vectors. The combined use of ATT&CK, ATHENA and ATLAS enables identification of plausible intrusion paths, escalation steps, and ML-targeted evasion techniques relevant to an in-vehicle CAN-IDS compromise.

During the "reconnaissance" phase, the adversary determines which ECU types are reachable (e.g., infotainment, telematics, CAN gateway). Of particular concern are wireless interfaces, used for passive anti-theft systems, Tire Pressure Monitoring Systems (TPMS), Bluetooth, FM data channels, and telematics [14]. Empirical studies demonstrate these pathways: gateway ECUs can be subverted to gain access to the CAN bus [14], and TPMS, Bluetooth, FM channels, and cellular networks have been compromised via reverse engineering and protocol abuse [14]. In addition, Bozdal et al. propose a remote vector involving a malicious self-diagnostic application that, once installed, enables long-range control and monitoring of a vehicle [14]. Notably, Valasek and Miller demonstrated remote compromise of a Jeep Cherokee, including engine disablement, in 2014 [14].

During the "weaponization" phase, the adversary develops or acquires the capabilities required to execute the intended intrusion. This may involve crafting payloads (for example, modified ECU firmware) or preparing mechanisms to inject CAN frames via over-the-air (OTA) update channels or continuous update processes [57, 58]. Supply-chain compromises that enable the physical replacement or insertion of malicious components remain among the most commonly observed strategies; for instance, Mahmood et al. demonstrated control of a Chevrolet Corvette's brakes and windshield wipers through a commercial telematics control unit in 2016 [58]. Social engineering techniques may also be

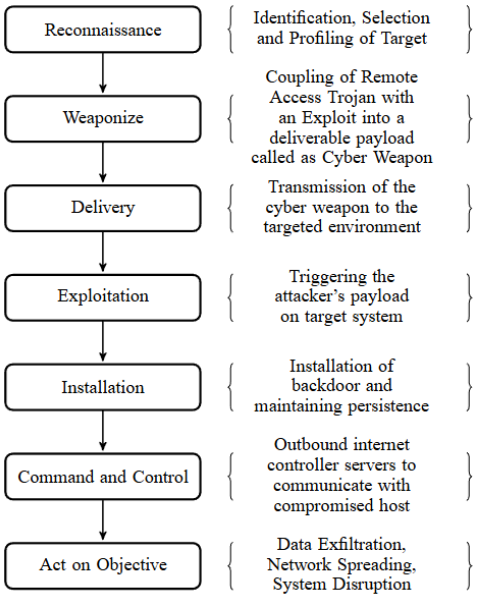


Figure 5: *Cyber* kill chain based on Mitre Att&ck and Atlas [56, 54]

employed to induce a vehicle operator to perform actions that facilitate exploitation (e.g., installing a compromised app or accepting an update). To the best of our knowledge, there is limited published evidence documenting scenarios in which such social-engineering vectors are combined with component substitution in operational attacks, although such combinations are not implausible.

During the "delivery" phase, the adversary executes the previously prepared vector to compromise the target ECU and trigger execution of malicious code (exploitation) [58]. The goal is to understand the topology of the vehicle network to understand where to move (towards the IDS). After gaining control of an ECU, the attacker may proceed, during the installation step, to subvert the in-vehicle IDS in order to degrade or neutralize detection capabilities. In this situation, the main assumption is the IDS module is direct connected to the CAN network. Following successful IDS compromise, the adversary can perform canonical CAN-bus operations, such as the injection of crafted malicious frames, and also to understand which attack conduct on the CAN network, then gather details on the information flow (which frames to check) through command-and-control channels. Finally, the direct manipulation of vehicle subsystems is performable [58, 59].

Following the Mitre Atlas knowledge base [56], during the "reconnaissance" phase, the adversary surveys the target landscape to identify susceptible manufacturers and vehicle models and to assess whether the onboard IDS module exhibits vulnerabilities to AML threats. This preliminary activity is reflected in the kill chain presented above and is consistent with reconnaissance methodologies described by Le et al. for identifying malicious URL-based attack surfaces [60].

In the "weaponization" phase, the attacker procures the capabilities required to carry out the intrusion, namely the necessary hardware components and the specific exploit implementations [61]. In the scenario examined in this work, those capabilities are deliberately limited to black-box techniques and very basic hardware components.

After infiltrating the CAN network and compromising an ECU, the adversary performs lateral movement to the IDS module and initiates an evasion attack by collecting labels associated with exfiltrated (preprocessed) CAN frames in order to obtain the adversarial one; these activities are conducted during the "command and control" stage. This adversarial approach parallels industry case studies: for example, the Palo Alto Networks AI Research Team demonstrate how oracle-style feedback can be exploited to evade a convolutional neural network-based DGA detector via controlled mutations of candidate domain names [62].

Finally, during the "action-on-objective" phase, the attacker submits adversarial examples to the

operational IDS, thereby undermining its integrity and producing downstream harms through the CAN attacks. The effects can be reputational damage or social disorder, outcomes analogous to those considered by Mithril Security Researchers in their analysis of poisoning attacks against large language models [63]. Figure 5 shows the previous kill chain steps [64]. To establish a proactive defensive posture and facilitate seamless symbiosis between the ML-based Intrusion Detection Systems (IDS) and human operators within a multidisciplinary Blue Team [65], this study evaluates the efficacy of hyperparameter optimization as an intrinsic defense mechanism to regulate Adversarial Machine Learning (AML) attack latencies. Furthermore, in alignment with the Framework to Create Symbiotic Artificial Intelligence (Frame SAI) [52], specifically Guideline 1.1 regarding human oversight—the victim IDS can be engineered to monitor query frequencies. This enables the prediction of critical threshold values, thereby empowering operators to supervise and intervene in the system's decisional logic.

5. Conclusion

The research work presented in this document explores the interplay between disinformation campaigns and cyber attacks against critical AI-enabled systems. The *Cyber* attack surfaces encompasses the interconnected electronic control units and ML-based IDS for in-vehicle automotive security.

Disinformation narratives are the core of the *informational* capabilities, while AML techniques are the central *Cyber* capability exploited. The *Cyber* component of the hybrid operation targets an IDS that monitors CAN traffic. Disinformation is not only an amplifier of the *Cyber* attack but also represents its starting point, preparing the victim community for an imminent threat. The goal of this "pre" *Cyber* attack phase is to create uncertainty and undermining the normal state of mind and sentiment of the community itself. The ultimate goal of the hypotized hybrid red team is to shape user perception and eroding public trust.

The research methodology started in reviewing adversarial tactics, particularly the historical pairing of disinformation and automotive *Cyber* attacks, and identifying this gap in current cyber threat intelligence. Building on this, a blue team process considering the ethical and legal aspects as central component useful to model the attack/defense pipeline is designed. An hybrid threat model mapped to an hybrid kill chain is examined to integrate the cognitive and technical dimensions of attack. The final objective is to create awareness about this types of threats and underline the need of possible countermeasures.

The hybrid threat model (Figure 2) posits that informational and *Cyber* sub-offensives operate in a semi-concurrent manner. An initial *informational* vector induces cognitive disruption by amplifying perceived vehicle malfunctions, after which a *Cyber* intrusion exploits technical vulnerabilities. Each sub-offensive reinforces the other, generating reciprocal amplification within the socio-technical environment. The corresponding hybrid kill chain (Figure 3) unfolds through three sequential flows: an *informational* offensive, a *Cyber* attack on the in-vehicle network, and a subsequent compromise of the IDS.

5.1. Future Work

From the perspective of the *Cyber* offense, the following future developments can be identified: implement a ML-based IDS considering state-of-the-art techniques for technology transfer, implement a system similar to a CAN network in order to study any lateral movements on the IDS, study the impact of state-of-the-art AML attacks taking into account multifaceted techniques (also from other data domains).

From the perspective of the hybrid offense, a future development can be the simulation of a disinformation attack leveraging a Multi-Agent system powered by Large Language Models (LLMs) to assess the effectiveness and feasibility of the proposed hybrid threat model. The simulation can involve the creation of a social media post for two distinct phases, namely the "pre" and "post" cyber attack, to fuel the disinformation campaign. By disseminating these posts on a simulated social media platform, it can then analyze the resulting user reactions and thus evaluate the hybrid attack's overall impact.

6. Acknowledgments

This work was partially supported by the following projects: SERICS - “Security and Rights In the CyberSpace - SERICS” (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU; CYBER-PREDICT: Cyber vulnerability ranking prediction by prescription, Avviso “Reti - Sostegno alla ricerca collaborativa”, - Codice Progetto DUQVKW0; Patto territoriale "Sistema universitario pugliese" – CUP F61B23000370006; "AURIGA-EDICO" (Ecological Digital Conversion), under the Regional Regulation n.17/2014 – PIA (art. 26) – Project code: 4WBHY73. The research was conducted in partnership with the Department of Computer Science, University of Bari Aldo Moro (SERLAB – Software Engineering Research Laboratory), focusing on the development of AI/ML-based tools and quality measurement systems for sustainable digital conversion of business processes; "CA-LCAP" (Low-Code Application Platform), under the Programma Operativo FESR 2014-2020 – Contratti di Programma (practice code 4UP72S1). The research was conducted in partnership with the Department of Computer Science, University of Bari, within the framework of a research and development contract (OR3 – Activities A.3.5 and A.3.6).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] M. Schulze, Cyber in war: Assessing the strategic, tactical, and operational utility of military cyber operations, in: 2020 12th International Conference on Cyber Conflict (CyCon), volume 1300, 2020, pp. 183–197. doi:10.23919/CyCon49761.2020.9131733.

[2] S. G. della Difesa Italiana, The italian defence approach to multi-domain operations (approccio della difesa alle operazioni multidominio), https://www.difesa.it/assets/allegati/31787/2.1defence_approach_to_mdos.pdf, 2022.

[3] N. A. T. O. (NATO), Nato 2022 strategic concept, 2022. URL: https://www.nato.int/cps/en/natohq/topics_210907.htm.

[4] E. Union, A strategic compass for security and defence, 2022. URL: https://www.eeas.europa.eu/eeas/strategic-compass-security-and-defence-0_en.

[5] K. Kruijver, N. B. Finlayson, B. Cadet, The disinformation lifecycle: an integrated understanding of its creation, spread and effects, Discover Global Society 3 (2024) 58. URL: https://doi.org/10.1007/s44282-025-00194-5. doi:10.1007/s44282-025-00194-5.

[6] E. Union, The code of conduct on disinformation, 2025. URL: https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation.

[7] L. E. R. Peters, G. E. C. Charnley, S. Roberts, I. Kelman, Public health disinformation, conflict, and disease outbreaks: a global narrative integrative review to guide new directions for health diplomacy, Global Health Action 18 (2025) 2562380. URL: https://doi.org/10.1080/16549716.2025.2562380. doi:10.1080/16549716.2025.2562380. arXiv:https://doi.org/10.1080/16549716.2025.2562380.

[8] S. R. Soare, J. Burton, Smart cities, cyber warfare and social disorder, 2020. URL: https://api.semanticscholar.org/CorpusID:235386282.

[9] M. T. Baldassarre, V. S. Barletta, D. Caivano, Smart program management in a smart city, in: 2018 AEIT International Annual Conference, 2018, pp. 1–6. doi:10.23919/AEIT.2018.8577379.

[10] T. Campisi, A. Severino, M. A. Al-Rashid, G. Pau, The development of the smart cities in the connected and autonomous vehicles (cavs) era: From mobility patterns to scaling in cities, Infrastructures 6 (2021). URL: https://www.mdpi.com/2412-3811/6/7/100. doi:10.3390/infrastructures6070100.

[11] S. K. Khan, N. Shiwakoti, P. Stasinopoulos, Y. Chen, M. Warren, Cybersecurity framework for connected and automated vehicles: A modelling perspective, Transport Policy 162 (2025) 47–64. URL: https://www.sciencedirect.com/science/article/pii/S0967070X24003561. doi:https://doi.org/10.1016/j.tranpol.2024.11.019.

[12] T. Junko, An overview of cyber security for connected vehicles, 2018. doi:10.1587/transinf.2017ICI0001.

[13] H. Olufowobi, G. Bloom, Chapter 16 - connected cars: Automotive cybersecurity and privacy for smart cities, in: D. B. Rawat, K. Z. Ghafoor (Eds.), Smart Cities Cybersecurity and Privacy, Elsevier, 2019, pp. 227–240. doi:https://doi.org/10.1016/B978-0-12-815032-0.00016-0.

[14] M. Bozdal, M. Samie, S. Aslam, I. Jennions, Evaluation of can bus security challenges, Sensors 20 (2020) 16–17. doi:10.3390/s20082364.

[15] A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman, Survey of intrusion detection systems: techniques, datasets and challenges, Cybersecurity 2 (2019). doi:10.1186/s42400-019-0038-7.

[16] J. Mehta, G. Richard, L. Lugosch, D. Yu, B. H. Meyer, Dt-ds: Can intrusion detection with decision tree ensembles, ACM Trans. Cyber-Phys. Syst. 7 (2023). URL: https://doi.org/10.1145/3566132. doi:10.1145/3566132.

[17] B. Wu, Z. Zhu, L. Liu, Q. Liu, Z. He, S. Lyu, Attacks in adversarial machine learning: A systematic survey from the life-cycle perspective, 2024. arXiv:2302.09457.

[18] V. S. Barletta, D. Caivano, C. Catalano, S. del Vescovo, M. Scalera, Towards the responsible/trustworthy ai in multi-domain operations for cyber social security: A black-box aml case study in the can bus frame detection task, volume 3962, 2025. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105005830931&partnerID=40&md5=517007d78702dcbabdba4d5e0650ef5b.

[19] M. Kalinowski, D. Mendez, G. Giray, A. P. Santos Alves, K. Azevedo, T. Escovedo, H. Villamizar, H. Lopes, T. Baldassarre, S. Wagner, S. Biffl, J. Musil, M. Felderer, N. Lavesson, T. Gorschek, Naming the pain in machine learning-enabled systems engineering, Information and Software Technology 187 (2025). doi:10.1016/j.infsof.2025.107866.

[20] N. A. T. O. (NATO), Hybrid threats and hybrid warfare reference curriculum, 2024. URL: https://www.nato.int/nato_static_fl2014/assets/pdf/2024/7/pdf/241007-hybrid-threats-and-hybrid-warfare.pdf.

[21] F.-S. Gady, A. Stronell, Cyber capabilities and multi-domain operations in future high-intensity warfare in 2030, Cyber Threats and NATO 2030: Horizon Scanning and Analysis (2020) 151.

[22] A. Gilli, M. Gilli, G. G. and, Nato, multi-domain operations and the future of the atlantic alliance, Comparative Strategy 44 (2025) 73–91. URL: https://doi.org/10.1080/01495933.2024.2445491. doi:10.1080/01495933.2024.2445491. arXiv:https://doi.org/10.1080/01495933.2024.2445491.

[23] Wikipedia, 2025 india–pakistan conflict, 2025. URL: https://en.wikipedia.org/wiki/2025_India%E2%80%93Pakistan_conflict.

[24] N. Y. Post, Eu leaders warn russia is waging a hybrid war on europe and must rearm: 'we have to be ready', 2025. URL: https://nypost.com/2025/10/01/world-news/russia-is-waging-a-hybrid-war-on-europe-and-must-rearm-denmark-warns/.

[25] K. Selvaraj, M. M. Singh, Z. F. Zaaba, A conceptual framework to mitigate ransomware attacks on iomt devices using threat intelligence: a systematic literature review, Computer Science Review 58 (2025) 100801. URL: https://www.sciencedirect.com/science/article/pii/S1574013725000772. doi:https://doi.org/10.1016/j.cosrev.2025.100801.

[26] E. I. Sharing, A. C. (E-ISAC), Analysis of the cyber attack on the ukrainian power grid, 2016. URL: https://nsarchive.gwu.edu/sites/default/files/documents/3891751/SANS-and-Electricity-Information-Sharing-and.pdf.

[27] BBC, How russian-funded fake news network aims to disrupt election in europe - bbc investigation, 2025. URL: https://www.bbc.com/news/articles/c4g5kl0n5d2o.

[28] L. Bidochko, Disinformation against crimean tatars in russian social media: Communication dynamics during the first year of russia's full-scale invasion of ukraine, Journal on Ethnopolitics and Minority Issues in Europe 23 (2024) 81–107. URL: https://www.ecmi.de/JEMIE/index.php/

journal/article/view/117. doi:10.53779/LEBI1430.
[29] A. Kruti, U. Butt, R. B. Sulaiman, A review of solarwinds attack on orion platform using persistent threat agents and techniques for gaining unauthorized access, 2023. URL: https://arxiv.org/abs/2308.10294. arXiv:2308.10294.
[30] B. Wu, M. Zhu, M. Zheng, Z. Zhu, S. Wei, M. Zhang, H. Chen, D. Yuan, L. Liu, Q. Liu, Defenses in adversarial machine learning: a systematic survey from the lifecycle perspective, IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 1–20. doi:10.1109/TPAMI.2025.3611340.
[31] G. Cloud, Sandworm disrupts power in ukraine using a novel attack against operational technology, 2023. URL: https://cloud.google.com/blog/topics/threat-intelligence/sandworm-disrupts-power-ukraine-operational-technology/.
[32] C. Wardle, H. Derakhshan, A. Burns, N. Dias, Information disorder: Toward an interdisciplinary framework for research and policy making (2017), 2017. URL: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html.
[33] F. S. González, J. Pastor-Galindo, J. A. Ruipérez-Valiente, Toward interoperable representation and sharing of disinformation incidents in cyber threat intelligence, 2025. URL: https://arxiv.org/abs/2502.20997. arXiv:2502.20997.
[34] T. Guardian, Russia using criminal networks to drive increase in sabotage acts, says europol, 2025. URL: https://www.theguardian.com/technology/2025/mar/18/russia-criminal-networks-drive-increase-sabotage-europol.
[35] C. Wardle, H. Derakhshan, Information disorder: Toward an interdisciplinary framework for research and policymaking, volume 27, Council of Europe Strasbourg, 2017.
[36] T. Moraca, F. Sicurella, T. Sekulic, J. Armanini, F. Stella, et al., Feeling the pulse. Countering foreign information manipulation and interference in Africa and the Western Balkans, JSTOR, 2023.
[37] Y. Danyk, C. M. Briggs, Modern cognitive operations and hybrid warfare, Journal of Strategic Security 16 (2023) 35–50.
[38] S.-D. Dov Bachmann, D. Putter, G. Duczynski, Hybrid warfare and disinformation: A ukraine war perspective, Global Policy 14 (2023) 858–869.
[39] M. A. Richter, M. Hagenmaier, O. Bandte, V. Parida, J. Wincent, Smart cities, urban mobility and autonomous vehicles: How different cities needs different sustainable investment strategies, Technological Forecasting and Social Change 184 (2022) 121857. URL: https://www.sciencedirect.com/science/article/pii/S004016252200381X. doi:https://doi.org/10.1016/j.techfore.2022.121857.
[40] E. Ismagilova, L. Hughes, N. P. Rana, Y. K. Dwivedi, Security, privacy and risks within smart cities: Literature review and development of a smart city interaction framework, Information Systems Frontiers 24 (2022). URL: https://doi.org/10.1007/s10796-020-10044-1. doi:10.1007/s10796-020-10044-1.
[41] X. Chen, Y. Deng, H. Ding, G. Qu, H. Zhang, P. Li, Y. Fang, Vehicle as a service (vaas): Leverage vehicles to build service networks and capabilities for smart cities, 2023. URL: https://arxiv.org/abs/2304.11397. arXiv:2304.11397.
[42] S. Rajapaksha, H. Kalutarage, M. O. Al-Kadri, A. Petrovski, G. Madzudzo, M. Cheah, Ai-based intrusion detection systems for in-vehicle networks: A survey, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3570954. doi:10.1145/3570954.
[43] A. Alshammari, M. Zohdy, D. Debnath, G. Corser, Classification approach for intrusion detection in vehicle systems, Wireless Engineering and Technology 09 (2018) 79–94. doi:10.4236/wet.2018.94007.
[44] A. Anjum, P. Agbaje, S. Hounsinou, H. Olufowobi, In-vehicle network anomaly detection using extreme gradient boosting machine, in: 2022 11th Mediterranean Conference on Embedded Computing (MECO), 2022, pp. 1–6. doi:10.1109/MECO55406.2022.9797224.
[45] I. Zenden, H. Wang, A. Iacovazzi, A. Vahidi, R. Blom, S. Raza, On the resilience of machine learning-based ids for automotive networks, in: 2023 IEEE Vehicular Networking Conference (VNC), IEEE, 2023. doi:10.1109/vnc57357.2023.10136285.

[46] K. T. Y. Mahima, M. Ayoob, G. Poravi, Adversarial attacks and defense technologies on autonomous vehicles: A review, Appl. Comput. Syst. 26 (2021) 96–106. doi:10.2478/acss-2021-0012.

[47] F. Aloraini, A. Javed, O. Rana, Adversarial attacks on intrusion detection systems in in-vehicle networks of connected and autonomous vehicles, Sensors 24 (2024). URL: https://www.mdpi.com/1424-8220/24/12/3848. doi:10.3390/s24123848.

[48] S. no To Disinformation, F. H. Communications, Can a.i. cause a bank run?, 2025. URL: https://www.saynotodisinfo.com/_files/ugd/438ee6_d9f4506bfd2e43218b96f716bae91ce1.pdf.

[49] S. Terp, P. Breuer, Disarm: a framework for analysis of disinformation campaigns, in: 2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), 2022, pp. 1–8. doi:10.1109/CogSIMA54611.2022.9830669.

[50] T. M. of Foreign Affairs of the Russian Federation, Examples of the large-scale disinformation campaign by the western media, hyping the claim that russia is planning to invade ukraine, 2022. URL: https://mid.ru/en/press_service/articles_and_rebuttals/rebuttals/nedostovernie-publikacii/1798160/.

[51] N. Capuano, G. Fenza, V. Loia, F. D. Nota, Content-based fake news detection with machine and deep learning: a systematic review, Neurocomputing 530 (2023) 91–103. URL: https://www.sciencedirect.com/science/article/pii/S0925231223001376. doi:https://doi.org/10.1016/j.neucom.2023.02.005.

[52] M. Calvano, A. Curci, R. Lanzilotti, A. Piccinno, Framesai: a three-layer framework to create symbiotic ai systems (2025). URL: https://api.semanticscholar.org/CorpusID:284593198.

[53] G. Desolda, A. Esposito, R. Lanzilotti, A. Piccinno, M. F. Costabile, From human-centered to symbiotic artificial intelligence: a focus on medical applications, Multimedia Tools and Applications 84 (2025) 32109–32150.

[54] B. Al-Sada, A. Sadighian, G. Oligeri, Mitre att&ck: State of the art and way forward, ACM Comput. Surv. 57 (2024). URL: https://doi.org/10.1145/3687300. doi:10.1145/3687300.

[55] C. Company, Athena (advanced threat knowledge-base for networked automotive), 2023. URL: https://athena.cycraft.ai/.

[56] J. Carroll, Strengthening ai critical infrastructure security with the mit ai risk repository and mitre atlas frameworks, European Conference on Cyber Warfare and Security 24 (2025) 71–79. doi:10.34190/eccws.24.1.3713.

[57] S. Mahmood, H. N. Nguyen, S. A. Shaikh, Systematic threat assessment and security testing of automotive over-the-air (ota) updates, Vehicular Communications 35 (2022) 100468. URL: https://www.sciencedirect.com/science/article/pii/S2214209622000158. doi:https://doi.org/10.1016/j.vehcom.2022.100468.

[58] N. Ayres, L. Deka, D. Paluszczyszyn, Continuous automotive software updates through container image layers, Electronics 10 (2021). URL: https://www.mdpi.com/2079-9292/10/6/739. doi:10.3390/electronics10060739.

[59] F. Sommer, J. Dürrwang, R. Kriesten, Survey and classification of automotive security attacks, Information 10 (2019). URL: https://www.mdpi.com/2078-2489/10/4/148. doi:10.3390/info10040148.

[60] H. Le, Q. Pham, D. Sahoo, S. C. H. Hoi, Urlnet: Learning a url representation with deep learning for malicious url detection, 2018. URL: https://arxiv.org/abs/1802.03162. arXiv:1802.03162.

[61] A. G. A. Team, Camera hijack attack on facial recognition system, 2020. URL: https://atlas.mitre.org/studies/AML.CS0004.

[62] P. A. N. A. R. Team, Botnet domain generation algorithm (dga) detection evasion, 2020. URL: https://atlas.mitre.org/studies/AML.CS0001.

[63] M. S. Researchers, Poisongpt, 2023. URL: https://atlas.mitre.org/studies/AML.CS0019.

[64] T. Yadav, A. M. Rao, Technical Aspects of Cyber Kill Chain, Springer International Publishing, 2015, p. 438–452. URL: http://dx.doi.org/10.1007/978-3-319-22915-7_40. doi:10.1007/978-3-319-22915-7_40.

[65] V. S. Barletta, D. Caivano, G. Cellammare, S. del Vescovo, A. L. Sciacovelli, Leveraging trustworthy ai for automotive security in multi-domain operations: Towards a responsive human-ai multi-domain task force for cyber social security, 2025. URL: https://arxiv.org/abs/2507.21145.

arXiv:2507.21145.