Malicious Reconfigurable Intelligent Surfaces: Security Threats in 6G Networks

Waqas Khalid, *Member, IEEE*, Trinh Van Chien, Wali Ullah Khan, *Member, IEEE*, Zeeshan Kaleem, *Senior Member, IEEE*, Yousaf Bin Zikria, *Senior Member, IEEE*, Taejoon Kim, *Member, IEEE*, and Heejung Yu, *Senior Member, IEEE*

***Abstract*—Reconfigurable intelligent surfaces (RISs) are emerging as a transformative technology for sixth-generation (6G) wireless networks. They enable dynamic manipulation of the propagation environment to enhance signal coverage, mitigate interference, and improve spectral and energy efficiencies. However, this flexibility introduces significant security vulnerabilities when RISs are maliciously controlled. This study explores the threats posed by such RISs, focusing on their potential to compromise the security and integrity of 6G networks. From an adversarial perspective, we analyze key attack vectors, including sophisticated jamming attacks that disrupt communication, eavesdropping attacks that intercept communications, and pilot contamination attacks that impair channel estimation accuracy, all contributing to severe performance degradation. For each attack, we detail the underlying mechanisms and adversarial optimization strategies designed to maximize impact. A case study quantifies the practical effects of these malicious RIS-based attacks in a simulated 6G network scenario. This research emphasizes the critical need for robust defense mechanisms and proposes essential research directions to address the evolving threats from malicious RISs, ensuring the security of 6G networks.**

***Index Terms*—Eavesdropping attacks, jamming attacks, malicious reconfigurable intelligent surface (RIS), pilot contamination attacks (PCA), sixth-generation (6G).**

Received 4 March 2025; revised 25 April 2025; accepted 9 May 2025. Date of publication 15 May 2025; date of current version 25 July 2025. This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT) under Grant RS-2025-00514779, and in part by the Korea University Grant. *(Corresponding authors: Heejung Yu; Taejoon Kim.)*

Waqas Khalid is with the Institute of Industrial Technology, Korea University, Sejong 30019, Republic of Korea (e-mail: waqas283@korea.ac.kr; waqas283@gmail.com).

Trinh Van Chien is with the School of Information and Communications Technology, Hanoi University of Science and Technology, Hanoi 100000, Vietnam (e-mail: chientv@soict.hust.edu.vn).

Wali Ullah Khan is with the Interdisciplinary Centre for Security, Reliability, and Trust, University of Luxembourg, 1855 Luxembourg City, Luxembourg (e-mail: waliullah.khan@uni.lu).

Zeeshan Kaleem is with the Department of Computer Engineering and Interdisciplinary Research Center for Smart Mobility and Logistics, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia (e-mail: zeeshan.kaleem@kfupm.edu.sa).

Yousaf Bin Zikria is with the Department of IT and Cyber Security, Institute of International Studies, Sydney, NSW 2000, Australia (e-mail: yousaf.zikria@tiis.edu.au).

Taejoon Kim is with the School of Information and Communication Engineering, Chungbuk National University, Cheongju 28644, Republic of Korea (e-mail: ktjcc@chungbuk.ac.kr).

Heejung Yu is with the Department of Electronics and Information Engineering, Korea University, Sejong 30019, Republic of Korea (e-mail: heejungyu@korea.ac.kr).

Digital Object Identifier 10.1109/JIOT.2025.3570544

I. Introduction

SIXTH-GENERATION (6G) wireless communication networks are expected to enable innovative use cases, exceptional services, and transformative applications by delivering ultrahigh transmission rates, ultralow latency, and unprecedented reliability [1], [2]. However, the inherent broadcast nature of the wireless medium, coupled with the escalating volume of sensitive data traversing these networks, including financial transactions, personal information, health records, and smart home automation commands, has elevated security to a paramount concern. While advancements in cryptographic techniques have strengthened upper-layer security, the physical layer remains vulnerable to sophisticated attacks [3]. Traditional physical-layer attacks, such as vulnerability analysis, information gathering, and forensic sniffing, are often detectable due to their reliance on leaving physical traces [4]. However, emerging technologies that enable precise manipulation of electromagnetic (EM) waves at the physical layer, particularly reconfigurable intelligent surfaces (RISs), have ushered in a new era of more insidious and challenging security threats. These novel attack vectors, leveraging the programmability of RISs, are significantly harder to detect and pose unprecedented challenges to the security and integrity of 6G networks [5], [6].

A. Vision for 6G: Smart Radio Environments Enabled by RISs

The vision for 6G networks extends beyond merely enhancing existing technologies; it aims to create truly smart radio environments enabled by RISs [7], [8], [9]. RIS technology marks a paradigm shift in wireless communication by transforming the traditionally static and uncontrollable wireless channel into a dynamic, programmable, and optimizable entity. By intelligently manipulating the reflection, refraction, and scattering of EM waves, RISs effectively mitigate common radio frequency (RF) impairments, including multipath fading, shadowing, and blockages. This capability significantly improves coverage, reliability, and overall network robustness, particularly in challenging environments [10]. Notably, RISs achieve these benefits with minimal complexity, requiring neither intricate signal processing nor power-hungry RF chains at the reflecting surface [11]. Traditionally, wireless system design has treated the radio environment as an immutable constraint, focusing optimization efforts solely

on the transmitter (Tx) and receiver (Rx) through techniques, such as multiantenna processing, advanced signal processing algorithms, and sophisticated communication protocols [12]. RIS technology fundamentally alters this approach by integrating the propagation environment itself into the optimization process. This allows for a holistic optimization of the entire communication link, including Tx, Rx, and channel, leading to substantial performance gains, including improved spectral and energy efficiency [13]. The strategic deployment and intelligent control of RISs, coupled with advanced beamforming techniques, unlock a range of benefits, such as extended signal coverage, reduced power consumption, increased throughput, effective interference mitigation, enhanced physical layer security (PLS), and more reliable communication links [14], [15], [16]. Thus, RISs are not merely an incremental improvement but a transformative technology poised to redefine the landscape of future wireless networks.

B. Security Perspective in 6G: The Rise of Physical Layer Security

As wireless communication networks evolve toward 6G, the need for robust security measures becomes increasingly critical. In this context, PLS has emerged as a powerful paradigm to complement traditional cryptographic methods and address the unique security challenges of the wireless domain [17], [18], [19], [20], [21], [22]. PLS exploits the inherent randomness and variability of the wireless channel to establish secure communication links, providing a dynamic and adaptive layer of defense. Unlike cryptography, which relies on computational complexity, PLS leverages the physical characteristics of the propagation environment, such as path loss, fading, noise, and interference, to enhance confidentiality and integrity [23]. A primary strength of PLS lies in its exploitation of channel disparities between legitimate users and eavesdroppers (Eves). Advanced techniques, such as directional beamforming focus signal energy toward legitimate users while forming spatial nulls to minimize leakage to Eves. Furthermore, PLS can employ artificial noise generation or targeted interference, strategically directed at Eve to disrupt interception [24], [25], [26]. In jamming scenarios, PLS techniques can mitigate the impact of malicious attacks by optimizing the signal-to-interference-plus-noise ratio (SINR) for legitimate users, ensuring reliable connectivity under hostile conditions [27].

C. Convergence of PLS and RIS in 6G: New Era of Wireless Security

RISs are poised to play a transformative role in strengthening PLS in 6G networks. This convergence of technologies represents a significant advancement in wireless security. By intelligently manipulating the reflection, refraction, and scattering of EM waves, RISs can dynamically reconfigure the wireless propagation environment, creating unique opportunities to enhance security [29]. For example, RISs can be configured to steer signals away from Eve, forming a secrecy-enhanced Area of Influence (AoI) around legitimate users [28].[1] They can also generate carefully controlled interference patterns to disrupt malicious jammers, effectively neutralizing their impact. By complementing traditional PLS techniques with the dynamic control offered by RISs, 6G networks can achieve a new level of adaptability and resilience against evolving security threats [30], [31], [32], [33]. A comprehensive analysis of these novel RIS-enhanced PLS techniques, highlighting their transformative potential for the 6G security paradigm, is presented in [34]. These techniques can proactively adapt to the presence of Eve or jammers, optimizing the communication environment in real time to maximize security.

D. Dark Side of RISs: Security Vulnerabilities in 6G

While RISs offer transformative advancements in 6G networks, particularly in enhancing security, their ability to dynamically control the wireless propagation environment also introduces significant security vulnerabilities. The inherent characteristics of RISs, such as their ability to manipulate signal reflection, refraction, and scattering, and the superposition of reflected signals, can be exploited by malicious actors. These capabilities, while beneficial for improving communication performance, simultaneously create novel attack vectors that are difficult to detect and counter. Most existing research focuses on leveraging RISs to enhance system performance, assuming they are controlled by legitimate network operators and meet security standards comparable to those of Tx and Rx. However, practical scenarios may reveal a different reality: illegally deployed or compromised RISs can be exploited to undermine network integrity and security [35]. Malicious actors could exploit RISs by hacking into their control systems, deploying unauthorized RISs with suboptimal or malicious configurations, or leveraging vulnerabilities inherent in low-cost RIS hardware. These actions pose substantial security threats to 6G networks. Unlike traditional active relays or large antenna arrays, RISs enable precise signal manipulation with minimal resources and low power consumption, offering adversaries a cost-effective means to execute sophisticated attacks [36]. This capability allows attackers to bypass conventional physical-layer defenses, paving the way for advanced physical-layer wireless attacks. The key security threats posed by malicious RISs are as follows.

1) *Eavesdropping:* An illegally controlled RIS can be manipulated to redirect and focus signals toward an Eve, enabling the adversary to intercept confidential information intended for the legitimate Rx [37]. This can be achieved by optimizing the RIS reflection coefficients to create constructive interference at Eve's location while potentially degrading legitimate Rx's signal.
2) *Jamming:* Malicious RIS can be configured to inject interference into the legitimate communication channel, degrading the signal quality at a legitimate Rx. This is achieved by manipulating the RIS elements to reflect

[1]The AoI denotes the spatial region where the RIS improves a QoS metric, such as secrecy spectral efficiency (SSE), thereby ensuring that the legitimate user's channel quality exceeds that of Eve and establishing a secure zone through physical-layer design.

signals that destructively interfere with the direct signal from the base station (BS), contrary to the performance enhancement achieved by legitimate RISs, thus disrupting network reliability and overall performance [38].

3) *Pilot Contamination:* During the channel estimation phase, malicious RIS can inject deceptive pilot signals or manipulate the reflection of legitimate pilots. This corrupts the channel state information (CSI) acquired by the BS, leading to suboptimal precoding and severe performance degradation for legitimate users [39].

Addressing these vulnerabilities is imperative as RIS deployment becomes increasingly integral to 6G networks. Mitigating these risks requires robust security frameworks, advanced anomaly detection mechanisms tailored to the unique characteristics of RIS technology, and potentially new PLS techniques that are resilient to RIS-aided attacks. A critical step toward securing RIS-aided communications involves aligning mitigation strategies with emerging standardization frameworks. Major bodies, such as the Third Generation Partnership Project (3GPP), the Institute of Electrical and Electronics Engineers (IEEE), and the International Telecommunication Union's Telecommunication Standardization Sector (ITU-T), actively integrate RISs into 6G architectures, leveraging the existing security mechanisms [40]. However, explicit guidelines for countering malicious RIS threats remain under development. Future standards are expected to introduce secure configuration protocols and defense mechanisms to neutralize the attack vectors associated with malicious RISs.

E. Motivation

While the vast majority of research on RISs focuses on their potential to enhance wireless network performance by mitigating threats, such as jamming and eavesdropping, or improving channel estimation accuracy, this article takes a fundamentally different perspective. We explore the emerging security risks posed by maliciously controlled RISs. These RISs, rather than being used to improve the network, can actively exploit vulnerabilities in the physical layer and wireless propagation environment to compromise the security and integrity of communication systems [41]. For instance, instead of minimizing signal leakage as in legitimate deployments, a malicious RIS can be strategically configured to facilitate eavesdropping by redirecting signals toward Eve. Similarly, a malicious RIS can be employed to reduce the SINR at legitimate Rxs, enabling effective intentional jamming attacks. Furthermore, such deployments can disrupt accurate channel estimation processes by causing pilot contamination, leading to severe degradation in network performance.

While RIS technology undoubtedly holds significant promise for enhancing the capabilities of 6G networks, its potential for misuse poses critical security challenges that must be addressed proactively. This article aims to shed light on the dark side of RIS technology, highlighting the urgent need for a security-centric paradigm shift in RIS-enabled wireless environments. We delve into specific attack strategies that can be employed using malicious RISs and emphasize the need to develop secure and resilient network architectures for 6G systems. This research underscores the importance of considering RISs not just as performance enhancers but also as potential threat surfaces in designing future wireless networks.

F. Contribution

This article provides a timely and significant contribution to wireless security by shifting the focus from the well-documented benefits of RIS technology to its potential for malicious exploitation. While the ability of RISs to enhance communication performance has been extensively validated, their vulnerability to misuse, particularly within the context of envisioned 6G use cases [42], introduces critical security challenges that remain largely unexplored [43]. This work fills this gap by making the following key contributions.

1) We identify and analyze key factors that make RIS systems vulnerable to unauthorized access, adversarial manipulation, and exploitative attacks. By examining potential threats, including malicious control of RIS behavior and disruption of EM wave propagation, we categorize attack types and assess their practical implications.
2) We provide an in-depth analysis of malicious attacks, focusing on novel strategies for jamming, eavesdropping, and pilot contamination. We dissect the underlying attack mechanisms and potential optimization strategies employed by adversaries, revealing the sophisticated and potentially devastating nature of these emerging threats.
3) We present a case study supported by numerical simulations to demonstrate the severe impact of malicious RIS on network performance and security.

These contributions lay the groundwork for future research on mitigating the security risks posed by malicious RISs and developing resilient defense strategies for 6G networks. While our primary focus is on threat modeling, game-theoretic frameworks, commonly used in legitimate RIS scenarios, offer a compelling extension. Zero-sum, nonzero-sum, and Stackelberg games effectively model strategic attacker-defender interactions under asymmetric information [44], while Bayesian learning manages high-dimensional uncertainty [45]. Future work may explore robust and adaptive game-theoretic formulations to enhance resilience against adversarial RIS behavior.

Fig. 1 illustrates the overall structure of this article, while a list of frequently used abbreviations is provided in Table I.

II. Factors Contributing to RIS Vulnerability

This section outlines potential vulnerabilities in the configuration and operation of RIS-aided networks, as illustrated in Fig. 2, which can be exploited by malicious actors to compromise the intended operation of the RIS and, consequently, the security and performance of the entire wireless network.

1) *Backhaul Link Vulnerabilities:* RISs typically rely on backhaul links to communicate with the core network and receive configuration instructions. These links can be vulnerable to interception, tampering, or disruption. Security vulnerabilities in the RIS microcontroller,

TABLE I
ACRONYMS AND ABBREVIATIONS

Acronym	Definition	Acronym	Definition
RIS	Reconfigurable intelligent surface	6G	Sixth-generation
EM	Electromagnetic	RF	Radio frequency
Tx	Transmitter	Rx	Receiver
PLS	Physical layer security	CSI	Channel state information
BS	Base station	FPGAs	Field-programmable gate arrays
DoS	Denial of service	AJ	Active jammer
CJ	Constant jammer	IJ	Intermittent jammer
RJ	Reactive jammer	AJ	Adaptive jammer
MISO	Multiple-input single-output	LS	Least square
ZF	Zero-forcing	AWGN	Additive white Gaussian noise
ND-RIS	Non-diagonal RIS	TDD	Time-division duplex
MIMO	Multiple-input multiple-output	BCD	Block coordinate descent
MINLP	Mixed-integer Non-linear program	SVD	Singular value decomposition
MRT	Maximal ratio transmission	SLNR	Signal-to-leakage-plus-noise ratio
SNR	Signal-to-noise ratio	PCA	Pilot contamination attack
DT	Data transmission	PT	Pilot transmission
Eve	Eavesdropper	DRL	Deep reinforcement learning
AoA	Angle-of-arrival	SDR	Semi-definite relaxation
SINR	Signal-to-interference-plus-noise ratio	ML	Machine learning
NOMA	Non-orthogonal multiple access	SDP	Semi-definite programming
EVD	Eigenvalue decomposition	RCG	Riemannian conjugate gradient

whether at the hardware or software level, can be exploited via these backhaul links. A security approach must encompass the entire protocol stack, from the physical layer (silicon-level hardware) up through the network functions, operating systems, and firmware running on the microcontroller. Misconfigurations or weaknesses in any of these layers could allow attackers to gain unauthorized access and manipulate the RIS for malicious purposes [46].

2) *Feedback Mechanism Exploitation:* In typical RIS-aided deployments, BS calculates the optimal reflection coefficients based on estimated CSI and transmits these coefficients to the RIS controller via a dedicated feedback link. While this mechanism is essential for dynamic network optimization, vulnerabilities in the feedback link could be exploited by adversaries. For example, an attacker could intercept or manipulate the transmitted CSI or the reflection coefficients themselves. This could allow them to influence the RIS configuration, degrading system performance or even redirecting signals toward Eve.
3) *Microcontroller Domain Security:* The RIS microcontroller is a critical component responsible for baseband signal processing, controlling the RIS elements, often using field-programmable gate arrays (FPGAs), and managing the RF front-end. Vulnerabilities in the microcontroller's hardware or software, such as insecure interfaces or exploitable backdoors, could enable unauthorized access. Such breaches could compromise the integrity of signal processing, disrupt intended RIS operations, and ultimately threaten the overall security of the network.
4) *RIS Element and Hardware Vulnerabilities:* A physical RIS panel, including the circuit board, copper backplane, and the individual elements with their phase-shift circuits, constitutes another potential attack surface. Physical tampering, damage, or manufacturing defects could compromise the intended operation of the RIS. Furthermore, malicious misconfigurations of the phase-shifting circuits could disrupt the designed signal reflection patterns, leading to degraded signal quality, compromised system reliability, and reduced network performance.
5) *EM Wave Manipulation:* The core functionality of RIS technology lies in its ability to precisely control EM wave propagation through mechanisms such as reflection, refraction, absorption, focusing, and polarization. While these capabilities are intended to enhance signal processing and network performance, they can be subverted. Malicious alterations in the RIS element configurations, particularly the phase-shift circuits, can disrupt the intended wave behavior. This can lead to signal distortion, degraded system performance, and increased vulnerability to eavesdropping or jamming attacks.

The large-scale deployment of RISs in 6G networks further amplifies security vulnerabilities because numerous reconfigurable elements expand the attack surface across backhaul links, control signaling, and synchronization processes. These complex deployments demand efficient, real-time solutions for anomaly detection, secure control, and adaptive defense mechanisms. Consequently, lightweight, distributed, and scalable security frameworks are essential to ensure robust protection without compromising performance. Additionally, dense RIS deployments by multiple operators in shared areas may result in unintentional reflections and interference, leading to signal leakage, security vulnerabilities, and throughput degradation. This highlights the need for coordinated, large-scale RIS optimization across coexisting infrastructures.

III. METHODS AND IMPACTS OF ATTACKS BY MALICIOUS RIS

This section assesses the feasibility and impacts of various attacks facilitated by malicious RISs. We analyze the

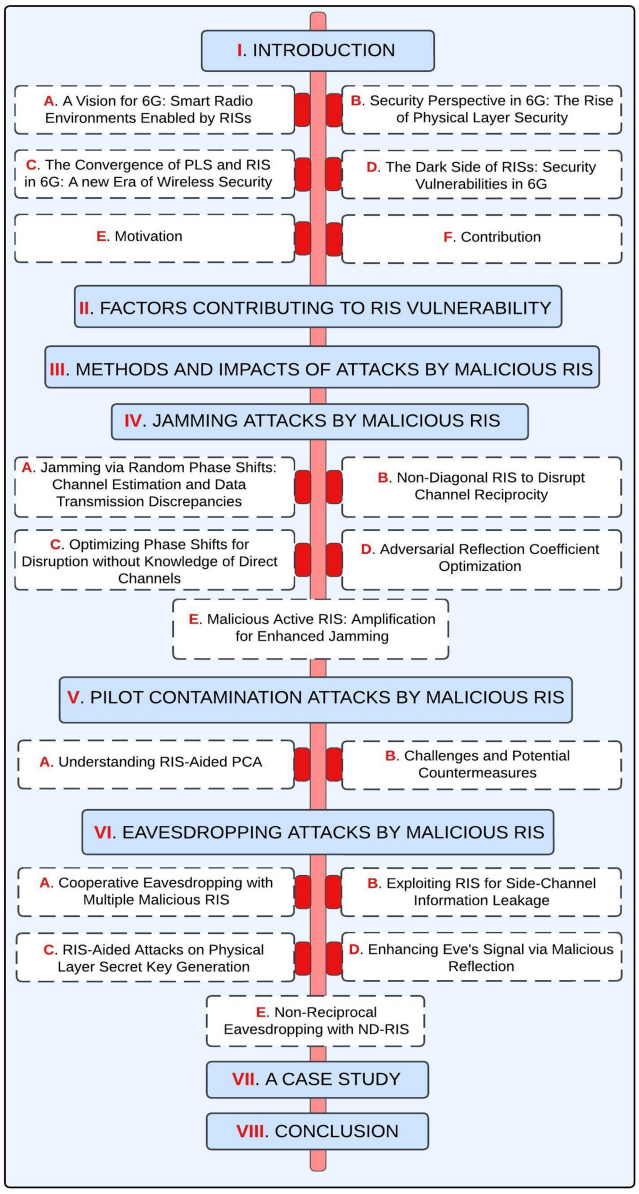


Fig. 1. Organization of this article.

sophisticated threat landscape, detailing attack types, wave manipulation mechanisms, employed strategies, prerequisites, and resulting performance degradation, as summarized in Table II. These attacks are categorized into two primary types: 1) jamming attacks (including SNR degradation at a legitimate Rx) and 2) eavesdropping attacks (including secrecy rate degradation).

A. Jamming Attacks

Adversaries manipulate an RIS reflection matrix (Φ) to introduce interference at a legitimate Rx. This can involve injecting random noise, creating nulls in the direction of Rx, exploiting nonorthogonal multiple access (NOMA) scheduling, using an active RIS for beamforming-based jamming, or configuring the RIS to absorb signals intended for the legitimate Rx. By leveraging the RIS configuration and, in some cases, CSI, attackers can optimize interference patterns, making these attacks both effective and difficult to detect [47].

B. Eavesdropping Attacks

Adversaries can exploit the RIS to redirect or amplify signals toward an Eve without necessarily degrading the signal quality for a legitimate Rx, though this may occur as a side effect. Precise control over the RIS's reflection matrix allows Eve to enhance the signal quality at Eve while minimizing any obvious signs of interception at the legitimate Rx. Using CSI, attackers can strategically position the RIS or manipulate its configuration to maximize intercepted information, compromising the confidentiality of the communication.

C. Advanced Attacks

A sophisticated class of attacks exploits nondiagonal RISs (ND-RISs), which possess nondiagonal reflection matrices unlike conventional RISs [48]. This capability allows ND-RISs to break channel reciprocity, creating discrepancies between uplink and downlink channels that can be exploited for eavesdropping and jamming attacks. An ND-RIS can remain passive during uplink channel estimation but apply a nondiagonal reflection matrix during the downlink, altering the channel to degrade the legitimate Rx's SNR while enhancing Eve's SNR. This mismatch renders the BS's uplink-based precoding suboptimal for the legitimate Rx. Additionally, an ND-RIS can generate strong jamming signals at a legitimate Rx, severely disrupting reception. Due to the ability of ND-RIS attacks to disrupt reciprocity, they are difficult to detect and require advanced hardware and control mechanisms [49].

This analysis reveals the sophisticated tactics attackers use to exploit RIS vulnerabilities. It underscores the need for robust countermeasures to secure 6G networks against the emerging threat of malicious RISs. The following sections will delve deeper into specific attack scenarios, providing further technical details and exploring potential mitigation strategies.

IV. Jamming Attacks by Malicious RIS

Jamming attacks, commonly classified as Denial of Service (DoS) attacks, are generally easier to execute than eavesdropping attacks due to their simpler operational requirements [50]. Traditionally, these attacks are initiated by an active jammer (AJ), which generates interference signals to disrupt communication between the BS and legitimate Rxs. Physical-layer AJs can be categorized into constant, intermittent, reactive, and adaptive types, as summarized in Table III [51], [52]. While conventional RIS technologies are designed to mitigate the effects of such external active jamming, a malicious RIS operates with the opposite intent: to amplify interference and degrade legitimate communications. The effectiveness of traditional AJs in reducing the SINR at legitimate Rxs depends on two key factors: 1) access to accurate CSI and 2) the ability to transmit active jamming signals. These prerequisites present notable challenges in practice. First, acquiring accurate CSI is inherently difficult due to the dynamic and distributed nature of wireless environments. Second, generating active

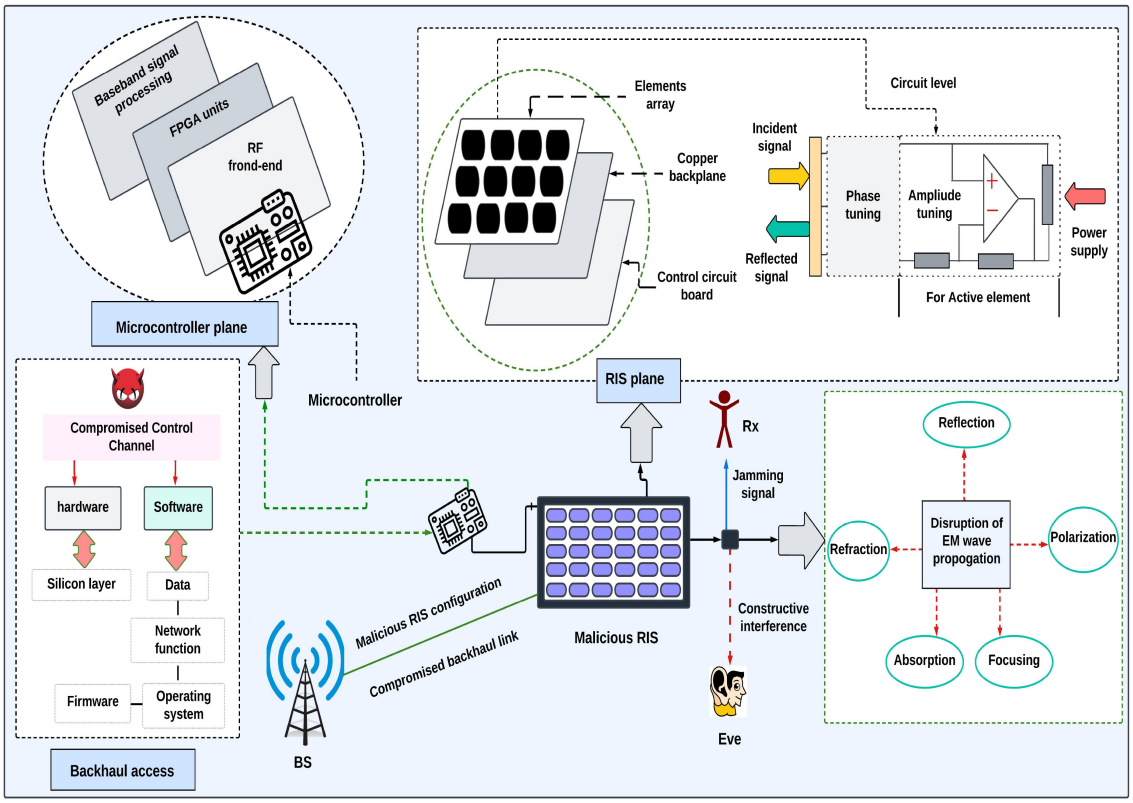


Fig. 2. Security vulnerabilities of malicious RISs.

jamming signals requires significant power resources, making such attacks potentially costly and easier to detect.

In this work, we focus on a more insidious threat: a malicious RIS acting as a passive jammer. This RIS does not actively transmit jamming signals but instead adjusts the amplitude and phase of the reflected signals to disrupt legitimate communications. Unlike AJ attacks, which are often detectable due to their active energy emissions, jamming by a malicious RIS is highly covert. The interference it introduces is often perceived as channel distortions or coverage issues, making it difficult to detect and mitigate using conventional methods [53]. In many proposed network designs, legitimate RISs are deployed near legitimate Rxs or BSs to optimize communication performance. In contrast, a malicious RIS, constrained by limited a priori information about user locations or precise CSI, is more likely to be strategically positioned closer to the BS to maximize its disruptive potential. By exploiting its proximity to the BS, the malicious RIS can significantly impact both uplink and downlink communications, particularly in scenarios where legitimate Rxs rely heavily on reflective channels for connectivity. This strategic positioning allows the malicious RIS to interfere with a larger number of legitimate Rxs or more critical communication links.

A. Jamming via Random Phase Shifts: Channel Estimation and Data Transmission Discrepancies

To investigate the innovative jamming strategies employed by a malicious RIS, we consider a multiuser multiple-input single-output (MISO) system.[2] Specifically, the system comprises a BS with N antennas communicating with K single-antenna legitimate users, i.e., legitimate Rxs, where $1 \leq k \leq K$. A malicious RIS with M reflecting elements is strategically positioned to disrupt the communication between the BS and legitimate users.

Given the data signal $s_i \in \mathbb{C}$ for the legitimate user i, normalized to unit power (i.e., $\mathbb{E}[|s_i|^2] = 1$), the signal received at the legitimate user k is expressed as

$$y_k = \mathbf{h}_{C,k}^H \sum_{i=1}^{K} \mathbf{w}_i s_i + n_k \tag{1}$$

where $\mathbf{h}_{C,k} = (\mathbf{H}_{B,R}^H \Phi \mathbf{h}_{R,k} + \mathbf{h}_{B,k}) \in \mathbb{C}^{N\times 1}$ represents the combined channel from the BS to user k, incorporating both the direct and RIS-cascaded channels. Specifically, $\mathbf{h}_{R,k} \in \mathbb{C}^{M\times 1}$ denotes a channel vector from the RIS to user k, $\mathbf{H}_{B,R} \in \mathbb{C}^{M\times N}$ is the channel matrix from the BS to the RIS and $\mathbf{h}_{B,k} \in \mathbb{C}^{N\times 1}$ is the channel vector from the BS to user k. The superscript H indicates the Hermitian transpose. The reflection matrix of the RIS, $\Phi = \text{diag}(\boldsymbol{\varphi}) \in \mathbb{C}^{M\times M}$, is defined in terms of the phase shift vector $\boldsymbol{\varphi} = [e^{j\varphi_1}, \ldots, e^{j\varphi_M}]^T$, where each phase shift φ_m is selected from the set $\Psi = \{0, 2\pi\}$ for $1 \leq m \leq M$ [54]. $\mathbf{w}_k$ represents the transmit beamforming vector for user k, and n_k denotes the additive white Gaussian noise (AWGN) at user k with a variance of σ^2.

[2]While this serves as the general system model under consideration, specific changes may be introduced depending on the jamming strategy.

TABLE II
THREAT LANDSCAPE POSED BY MALICIOUS RISS: ATTACK TYPES, MECHANISMS, STRATEGIES, PREREQUISITES, AND IMPACT

Attack type	Propagation mechanism	Malicious strategy	Attack prerequisites	Performance degradation/impact
Jamming / degrading SNR	Reflection with random noise	Inject random Gaussian phase offsets into RIS elements via the reflection matrix, introducing artificial interference patterns at the legitimate Rx	• Knowledge of RIS reflection matrix ($\mathbf{\Phi}$)	• Reduction in average SNR at a legitimate Rx. • Effectiveness depends on randomness and power of injected noise.
	Reflection with null-steering	Design a reflection matrix ($\mathbf{\Phi}$) that creates a null toward a legitimate Rx, minimizing its received signal power. This involves making $\mathbf{\Phi}$ orthogonal to the legitimate channel or exploiting the legitimate Rx's null space.	• Knowledge of RIS reflection matrix ($\mathbf{\Phi}$) • CSI of BS-RIS ($\mathbf{H}_{B,R}$) and RIS-Rx ($\mathbf{h}_R$) channels.	• SINR degradation at a legitimate Rx. • Compromising the reliability and quality of legitimate communications.
	Refraction / splitting	Exploit the reflection matrix and NOMA multi-user scheduling to manipulate user ordering and power allocation, degrading specific user performance	• CSI of all channels • Knowledge of RIS reflection matrix ($\mathbf{\Phi}$), NOMA scheduling, and power allocation schemes.	• Disruption of user scheduling and power balancing in NOMA. • Performance degradation for targeted users (e.g., Rx).
	Beamforming-based jamming (Active RIS)	Utilize an active RIS with controllable amplification to focus jamming signals toward Rx. Requires precise control over both phase and amplitude of reflected signals.	• Knowledge of RIS reflection matrix ($\mathbf{\Phi}$), including amplification factors • CSI of all channels. • Active RIS hardware.	• Substantial reduction in the legitimate Rx's SNR. • Potentially causing connectivity disruptions and degraded QoS. • Can be highly effective if RIS can steer a strong jamming beam.
	Absorption / blocking	Configure RIS elements to absorb incident signals, minimizing reflections toward a legitimate Rx. This can involve setting phase shifts to values that cause destructive interference or using materials that absorb RF energy.	• Knowledge of RIS reflection matrix ($\mathbf{\Phi}$). • May require specific RIS hardware capabilities.	• Near-total attenuation of reflected signals toward Rx. • Effectiveness depends on the RIS's ability to absorb energy and the relative strength of the direct path.
	Non-reciprocal jamming (ND-RIS)	Employ an ND-RIS to break channel reciprocity by remaining passive during uplink training and applying a non-diagonal reflection matrix during the downlink, generating strong jamming signals at Rx	• Control over an ND-RIS • CSI of relevant channels • Knowledge of a BS precoding scheme. • Requires sophisticated hardware for non-diagonal reflection.	• Substantial degradation of the legitimate Rx's SNR due to suboptimal precoding at the BS. • Difficult to detect due to passive uplink operation.
Eavesdropping / secrecy rate degradation	Reflection-based Eavesdropping	Adjust the RIS reflection matrix ($\mathbf{\Phi}$) to redirect signals toward Eve, optimizing Eve's SNR while potentially degrading the legitimate Rx's signal quality.	• Knowledge of RIS reflection matrix ($\mathbf{\Phi}$) • CSI of all channels, including BS-Eve ($\mathbf{h}_{B,E}$), RIS-Eve ($\mathbf{h}_{R,E}$) channels.	• Enhanced SNR at Eve. • Compromising secrecy and potentially enabling successful interception of sensitive data. • Reduced SNR at a legitimate Rx.
	Beamforming for eavesdropping	Leverage CSI and reflection matrix information to configure the RIS to act as a beamformer, focusing the reflected signal toward an Eve for optimized signal interception.	• Knowledge of RIS reflection matrix ($\mathbf{\Phi}$) • CSI of all channels, including BS-Eve ($\mathbf{h}_{B,E}$) and RIS-Eve ($\mathbf{h}_{R,E}$) channels.	• Significant reduction in secrecy capacity. • Eve's SNR is prioritized over the legitimate Rx's SNR, potentially leading to successful eavesdropping.
	Non-reciprocal eavesdropping (ND-RIS)	Employ an ND-RIS to break channel reciprocity by remaining passive during uplink training and applying a non-diagonal reflection matrix during the downlink, enhancing Eve's SNR and degrading the legitimate Rx's SNR.	• Control over an ND-RIS • CSI of relevant channels • Knowledge of BS precoding schemes. • Requires sophisticated hardware for non-diagonal reflection.	• Substantial degradation of the legitimate Rx's SNR due to suboptimal precoding at the BS. • Enhanced Eve's SNR. • Difficult to detect due to passive uplink behavior.

In standard RIS-assisted systems, the primary objective is to optimize key performance metrics, such as the achievable sum rate, which is defined as

$$R_T = \sum_{k=1}^{K} \log_2\left(1 + \frac{|\mathbf{h}_{C,k}^H \mathbf{w}_k|^2}{\sum_{i=1, i\neq k}^{K} |\mathbf{h}_{C,k}^H \mathbf{w}_i|^2 + \sigma^2}\right) \quad (2)$$

To maximize (2), the BS is required to compute the multiuser active beamforming $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K]$ that maximizes the desired signal power for the users (e.g., $|\boldsymbol{h}_{C,k}^H \mathbf{w}_k|^2$ for user k) while minimizing interuser interference (e.g., $\sum_{i=1, i\neq k}^{K} |\boldsymbol{h}_{C,k}^H \mathbf{w}_i|^2$ for user k). To optimize $\mathbf{W}$, the BS estimates the combined channel matrix $\mathbf{H}_C = [\mathbf{h}_{C,1}, \mathbf{h}_{C,2}, \ldots, \mathbf{h}_{C,K}]^H$, which captures the aggregated channel for all users. The channel estimation is typically performed during the uplink pilot phase using channel estimation algorithms such as the least squares (LS) approach [55]. The malicious RIS prevents the BS from obtaining separate CSI for the direct and RIS-assisted channels because the composite

TABLE III
CLASSIFICATION OF AJS AND THEIR STRATEGIC INSIGHTS

Jammer type	Ref.	Jamming mechanism	Strategic insights
Constant jammer (CJ)	[51]	Continuously transmit jamming signals, typically using pseudo-random noise or Gaussian-modulated waveforms.	• High detectability due to constant energy emissions and persistent interference patterns. • Inefficient in energy usage, suitable for basic DoS scenarios where stealth is unnecessary. • Limited adaptability to changing communication conditions.
Intermittent jammer (IJ)	[51]	Transmit jamming signals periodically, alternating between active and inactive states.	• Improved energy efficiency compared to CJs but vulnerable to exploitation during inactive periods. • Lower risk of detection due to non-continuous transmission. • Requires precise timing to disrupt communication during critical signal exchanges.
Reactive jammer (RJ)	[52]	Activate jamming signals only when legitimate communication is detected in the channel.	• High precision in actively disrupting communications through targeted jamming, minimizing unnecessary energy usage. • Complex implementation due to the need for real-time detection and fast response mechanisms. • Lower detectability compared to CJs and IJs due to selective interference.
Adaptive jammer (AJ)	[52]	Dynamically adjust jamming power and timing based on real-time wireless channel conditions and CSI.	• Optimizes jamming efficiency by exploiting channel state variations, ensuring minimal energy wastage. • High complexity, requiring continuous updates of CSI and advanced decision-making algorithms. • Suited for advanced theoretical models but challenging for real-world implementation due to computational overhead.

nature of the received signal hinders individual channel estimation. However, the BS can estimate the combined channels $\mathbf{h}_{C,k}$ once the malicious RIS fixes its reflection matrix. Using this estimate, the BS typically implements zero-forcing (ZF) beamforming to project each legitimate user's channel $\mathbf{h}_{C,k}$ onto a subspace orthogonal to the channel vectors of other legitimate users, thereby eliminating interuser interference.

1) Jamming Strategy by Malicious RIS With CSI: In an idealized scenario, the malicious RIS possesses complete CSI for all channels, enabling it to disrupt legitimate communications without transmitting explicit jamming signals. The BS, relying solely on the direct channel CSI, computes the multiuser active beamforming matrix $\mathbf{W}_d$ using ZF, where the subscript d indicates that the beamforming is based exclusively on the CSI of the direct channel. With full knowledge of $\mathbf{H}_{B,R}$, $\mathbf{h}_{R,k}$, and $\mathbf{h}_{B,k}$, along with the BS's designed $\mathbf{W}_d$, the malicious RIS strategically optimizes its reflection phase vector to minimize the sum rate of legitimate users. The optimization problem is formulated as

$$\min_{\boldsymbol{\varphi}} \sum_{k=1}^{K} \log_2\left(1+\frac{|\mathbf{h}_{C,k}^H \mathbf{w}_{d,k}|^2}{\sum_{i=1,i\neq k}^{K}|\mathbf{h}_{C,k}^H \mathbf{w}_{d,i}|^2+\sigma^2}\right) \tag{3a}$$

$$\text{s.t.}\quad \varphi_m \in \Psi, \quad m=1,2,\ldots,M \tag{3b}$$

where each phase shift φ_m is selected from the discrete set: $\Psi=\{0,(2\pi/L),\ldots,(2\pi(L-1)/L)\}$, with $L=2^b$, where b denotes the bit resolution of the phase shifters.

The optimization problem (3) involves enumerating all possible combinations of $\{\varphi_m\}_{m=1}^{M}$, resulting in 2^M discrete configurations and leading to significant computational complexity. To address this, the discrete phase shift constraint is relaxed to a continuous domain. Specifically, the phase shift vector is redefined as $\hat{\boldsymbol{\varphi}}=[e^{j\hat{\varphi}_1},\ldots,e^{j\hat{\varphi}_M}]^T$, allowing the phase shifts to vary continuously. This relaxation enables the reformulation of the optimization problem as

$$\min_{\hat{\boldsymbol{\varphi}}} \sum_{k=1}^{K} \log_2\left(1+\frac{|(\mathbf{h}_{R,k}\text{diag}(\hat{\boldsymbol{\varphi}})\mathbf{H}_{B,R}^H+\mathbf{h}_{B,k})\mathbf{w}_{d,k}|^2}{\sum_{i=1,i\neq k}^{K}|(\mathbf{h}_{R,k}\text{diag}(\hat{\boldsymbol{\varphi}})\mathbf{H}_{B,R}^H+\mathbf{h}_{B,k})\mathbf{w}_{d,i}|^2+\sigma^2}\right) \tag{4a}$$

$$\text{s.t.}\quad 0\le\hat{\varphi}_m\le 2\pi, \quad m=1,2,\ldots,M. \tag{4b}$$

This continuous formulation enables the use of efficient optimization techniques, such as the Riemannian conjugate gradient (RCG) algorithm. The constraint $\hat{\varphi}_m\in[0,2\pi]$ defines a complex circle manifold, making RCG an appropriate choice as it optimizes functions on Riemannian manifolds while ensuring feasibility. After obtaining the continuous solution $\hat{\boldsymbol{\varphi}}$, a nearest-neighbor search identifies the discrete phase shift vector closest to $\hat{\boldsymbol{\varphi}}$ in Euclidean distance, aligning the solution with the RIS hardware constraints.

Robust Optimization With CSI Uncertainty: The optimization problem (3) assumes perfect CSI at the malicious RIS. To incorporate CSI uncertainty, the problem can be reformulated as a min–max optimization that minimizes the sum rate under worst-case channel estimation errors

$$\min_{\boldsymbol{\varphi}} \max_{\|\Delta\mathbf{h}_{C,k}\|_2\le\epsilon_k} \sum_{k=1}^{K}\log_2\left(1+\frac{\left|\left(\widehat{\mathbf{h}}_{C,k}+\Delta\mathbf{h}_{C,k}\right)^H\mathbf{w}_{d,k}\right|^2}{\sum_{\substack{i=1\\i\neq k}}^{K}\left|\left(\widehat{\mathbf{h}}_{C,k}+\Delta\mathbf{h}_{C,k}\right)^H\mathbf{w}_{d,i}\right|^2+\sigma^2}\right) \tag{5a}$$

$$\text{s.t.}\quad \varphi_m\in\Psi \quad \forall m=1,2,\ldots,M \tag{5b}$$

where $\mathbf{h}_{C,k}=\widehat{\mathbf{h}}_{C,k}+\Delta\mathbf{h}_{C,k}$, and $\|\Delta\mathbf{h}_{C,k}\|_2\le\epsilon_k$ defines a bounded uncertainty. $\widehat{\mathbf{h}}_{C,k}$ and $\Delta\mathbf{h}_{C,k}$ denote an estimated channel and its estimation error.

This robust formulation minimizes the sum rate under the worst-case estimation error. Solving it typically involves applying the S-procedure to handle the inner maximization, followed by semidefinite relaxation (SDR) or block coordinate descent (BCD) to address the nonconvex phase constraints [56]. These techniques transform the problem into a more manageable form, though they may introduce approximation errors or require iterative refinement.

Remarks: Despite the effectiveness of CSI-dependent jamming strategy in disrupting legitimate communications, obtaining CSI for all channels and the BS's beamforming vectors remains a significant practical challenge for the malicious RIS. The robust formulation in (5) addresses CSI uncertainty and provides a foundation for evaluating the performance of malicious RIS under imperfect information. However, extending this robust approach to all attack scenarios is beyond the current scope and is left for future work.

2) Jamming Strategy by Malicious RIS Without CSI: A malicious RIS can operate without CSI by exploiting active channel aging, a method that creates temporal inconsistencies in the effective channel. This manipulation disrupts the orthogonality between the multiuser active beamforming vectors and the co-user channels, thereby increasing interuser interference [57]. The attack proceeds as follows.

1) *Pilot Transmission Phase and Beamforming Computation:* During the pilot transmission phase, the malicious RIS sets its reflecting phase vector to $\boldsymbol{\varphi}^a$, where the phase of each element φ_m^a is randomly assigned a phase shift from a uniform distribution over the set Ψ, i.e., $\varphi_m^a \sim \mathcal{U}(\Psi)$. This randomization perturbs the estimated channel vectors, resulting in $\mathbf{h}_{C,k}^a$ for $k = 1, \ldots, K$. Based on the perturbed channel estimate $\mathbf{h}_{C,k}^a$, the BS computes the multiuser active beamforming vector $\mathbf{w}_k^a$, ensuring orthogonality to the estimated co-user channels, i.e., $(\mathbf{h}_{C,i}^a)^H w_k^a = 0$ for $i = 1, \ldots, k-1, k+1, \ldots, K$.
2) *Data Transmission Phase:* During the data transmission phase, the malicious RIS modifies its reflecting phase vector to a different configuration, $\boldsymbol{\varphi}^b$. This unexpected modification disrupts the previously computed beamforming vector $\mathbf{w}_k^a$, causing it to lose orthogonality with the actual co-user channels during data transmission, i.e., $(\mathbf{h}_{C,i}^b)^H w_k^a \neq 0$ for $i = 1, \ldots, k-1, k+1, \ldots, K$.

Remarks: The malicious RIS introduces random phase shifts without CSI, causing the computed active beamforming vectors to be misaligned with the actual user channels during data transmission. This misalignment induces *active channel aging*, significantly increasing interuser interference and effectively jamming legitimate communications. The attack's effectiveness stems from the temporal inconsistency created between the channel estimation and data transmission phases.

B. Nondiagonal RIS to Disrupt Channel Reciprocity

In the previous section, we examined a jamming strategy where a malicious RIS introduces random phase shifts to disrupt a multiuser MISO system. While effective, this approach requires the malicious RIS to be synchronized with the legitimate system's timing and may be detectable due to rapid CSI variations that could exceed the channel coherence time. In this section, we introduce a more sophisticated and covert attack strategy leveraging an ND-RIS. Unlike conventional RISs, which employ diagonal reflection matrices, an ND-RIS utilizes a nondiagonal, nonsymmetric reflection matrix [58]. This fundamentally transforms the signal reflection mechanism. Unlike conventional RISs, where each element independently applies a phase shift and amplitude scaling to the incident signal, an ND-RIS enables interelement coupling. Specifically, the reflected signal at each element is a weighted linear combination of the signals impinging on all elements, with tunable phase shifts and potential amplitude adjustments incorporated during interelement signal exchange. This capability allows the ND-RIS to induce nonreciprocal signal propagation in a time-division duplex (TDD)-based system, thereby effectively decoupling the uplink and downlink channels.

1) Attack Mechanism: The ND-RIS remains passive during the uplink channel estimation phase, allowing the BS to estimate the uplink channels based on pilot signals transmitted by legitimate users, as in standard TDD operations. However, during the downlink transmission phase, the ND-RIS activates its nondiagonal reflection matrix. This introduces nonreciprocal phase shifts, effectively altering the downlink channels compared to the uplink ones. Since the BS relies on the uplink channel estimates (assuming reciprocity), it computes beamforming vectors that become nonoptimal for the actual downlink channels. This discrepancy leads to a significant increase in multiuser interference, thereby degrading system performance, particularly the SNRs experienced by legitimate users. Even if a diagonal reflection matrix is used during the downlink phase, differing from that in the uplink, channel reciprocity can still be broken, making the attack effective and difficult to detect.

2) Advantages of the ND-RIS Strategy: The ND-RIS offers enhanced stealthiness through the following attributes:

1) Unlike actively transmitting jammers or RISs that inject random phase shifts during channel estimation, the ND-RIS remains passive during the uplink, making it harder to detect by conventional detection mechanisms.
2) ND-RIS operates asynchronously with the legitimate system, requiring no temporal or frequency synchronization for channel estimation or data transmission, allowing it to bypass synchronization-based countermeasures.
3) The attack does not necessarily require explicit CSI of legitimate users. Instead, it exploits fundamental reciprocity assumptions in TDD systems.
4) Unlike other attacks that introduce rapid CSI variations, the ND-RIS does not artificially shorten the channel coherence time, further enhancing its stealthiness.

Remarks: These attacks significantly degrade the performance of legitimate users due to increased multiuser interference resulting from the BS's nonoptimal beamforming, which may be misattributed to increased noise or channel fading. While not strictly required, acquiring CSI enables the attacker to further optimize the ND-RIS reflection matrix,

intensifying the attack's impact. The ND-RIS-based attacks represent a major security threat to future wireless networks due to their ability to disrupt channel reciprocity passively without requiring synchronization or, initially, CSI.

C. Optimizing Phase Shifts for Disruption Without Knowledge of Direct Channels

In this section, we examine a more advanced attack strategy in which a malicious RIS optimizes its phase shifts to maximize disruption, even without direct knowledge of the BS-Rx channels. Under normal operations, the BS uses precise CSI to design beamforming vectors that mitigate interuser interference and enhance spatial multiplexing. However, we assume the malicious RIS has access to partial channel knowledge, specifically, the CSI of its cascaded channels, while remaining unaware of the direct BS-Rx channels. Using this knowledge, the attacker can manipulate impinging signals from the BS and passively reintroduce interuser interference that traditional precoding techniques cannot mitigate.

1) Beamforming at the BS (Without RIS-Induced Disruption): During the channel estimation phase, when the malicious RIS is inactive or operating in a way that does not significantly alter the channel, the BS computes beamforming vectors based on the estimated direct BS-Rx CSI. The vector $\mathbf{w}_k$ is typically designed to have two components: one is to mitigate interuser interference and the other is to enhance the signal reception at the Rx, i.e., user k. However, in an MISO system transmitting a single spatial stream, the second component is not defined. Please note that the second component to enhance the signal reception after nulling the interuser interference will be considered only when multiple symbol streams are transmitted with MIMO configurations. A common approach for designing $\mathbf{w}_k$ in the MISO case is outlined as follows.

1) *Covariance Matrix and Eigenvalue Decomposition (EVD):* The covariance matrix of the estimated direct channel for user k is calculated as $\Gamma_k = \mathbb{E}\{\mathbf{h}_{B,k}\mathbf{h}_{B,k}^H\} = \mathbf{V}_k\Lambda_k\mathbf{V}_k^H$, where $\Lambda_k \in \mathbb{R}^{r_k\times r_k}$ is a diagonal matrix containing the r_k largest eigenvalues of Γ_k (with r_k denoting the rank of Γ_k), and $\mathbf{V}_k \in \mathbb{C}^{N\times r_k}$ contains the corresponding eigenvectors.
2) *Channel Reformulation:* The estimated channel vector can be reformulated as $\mathbf{h}_{B,k} = \mathbf{V}_k\Lambda_k^{1/2}\tilde{\mathbf{h}}_{B,k}$, where $\tilde{\mathbf{h}}_{B,k} \in \mathbb{C}^{r_k\times 1}$ is a lower-dimensional abstraction of the channel, capturing the dominant fast-fading components.
3) *Interference Suppression Matrix:* To suppress interuser interference, the orthogonality condition $[\mathbf{V}_1\ldots\mathbf{V}_{k-1}\mathbf{V}_{k+1}\ldots\mathbf{V}_K]^H\mathbf{f}_k = \mathbf{B}_k^H\mathbf{f}_k = \mathbf{0}$ must hold, where $\mathbf{B}_k$ is a block matrix containing the eigenvectors corresponding to the largest eigenvalues of the nonintended users. The vector $\mathbf{f}_k$ is derived from the orthogonal complement of the column space of $\mathbf{B}_k$, typically using singular value decomposition (SVD) or other suitable matrix factorization techniques.

This hierarchical precoding strategy effectively suppresses interuser interference and, if an MIMO configuration is considered, optimizes signal reception in the absence of a malicious RIS.[3] However, a malicious RIS can severely degrade beamformer performance by leveraging even partial CSI.

2) Jamming Strategy by Malicious RIS: With the proposed beamforming design, the BS suppresses interuser interference in the direct channel by enforcing the orthogonality condition $\mathbf{h}_{B,i}^H\mathbf{w}_k = 0$ for $i \neq k$. However, the malicious RIS reintroduces interference through the RIS-reflected path. The signal received at user k is expressed as

$$y_k = \underbrace{\mathbf{h}_{B,k}^H\mathbf{w}_k x_k}_{\text{Desired Signal}} + \underbrace{\left(\mathbf{H}_{B,R}^H\Phi\mathbf{h}_{R,k}\right)^H\sum_{i=1}^{K}\mathbf{w}_i x_i}_{\text{RIS-Induced Interference}} + n_k \tag{6}$$

where the RIS-induced interference term introduces a nonnegligible cross-user interference that the BS cannot mitigate.

The malicious RIS leverages the knowledge of its cascaded channels ($\mathbf{H}_{B,R}$ and $\mathbf{h}_{R,k}$) to optimize its phase shifts Φ for disruption, despite lacking the CSI of direct BS-Rx channel ($\mathbf{h}_{B,k}$) or knowledge of the BS's beamforming vector ($\mathbf{w}_i$, for $i = 1,\ldots,K$). Although the BS employs hierarchical precoding to suppress direct cross-user interference, i.e., $\sum_{i\neq k}\mathbf{h}_{B,i}^H\mathbf{w}_i x_i$, it cannot counteract the RIS-induced interference term, i.e., $\sum_{i=1}^{K}(\mathbf{H}_{B,R}^H\Phi\mathbf{h}_{R,k})^H\mathbf{w}_i x_i$. This term remains an uncontrolled perturbation to the desired signal, making even the $i = k$ component a contributor to interference. The malicious RIS optimizes Φ to maximize the total reflected power, exacerbating interference for all users regardless of whether $i = k$ or $i \neq k$. The RIS-induced perturbation remains unstructured from the BS's perspective, rendering it unable to enforce constructive or destructive alignment. While, in certain cases, the RIS-reflected signal may inadvertently align constructively with the desired signal at some users, the overall statistical dominance of cross-user interference ensures a net SINR degradation. Since the interference scales with K, even sporadic enhancements of the desired signal are overshadowed by the amplified interference from other users. Consequently, SINR degradation persists, reaffirming the disruptive effect of the malicious RIS.

Remarks: A malicious RIS can implement a flexible jamming framework by dynamically adjusting the attack intensity for each user. This can be achieved by nonuniform weighting coefficients $\{\alpha_k\}$ to the reflected power toward each user k. For instance, high-priority users (e.g., those handling latency-sensitive or security-critical traffic) can be targeted with larger α_k, intensifying the impact of interference. However, optimizing the RIS-induced interference poses a nonconvex problem due to the unit-modulus constraints on the RIS phase shifts. To solve this problem efficiently, suboptimal solutions can be obtained through element-wise gradient ascent on the Riemannian manifold of phase shifts, SDR with randomization for feasible solutions, and deep reinforcement learning (DRL) for adaptive strategies in dynamic environments.

[3]If the system model is extended to an MIMO configuration, the overall precoding vector for user k is defined as $\mathbf{W}_k = \mathbf{F}_k\mathbf{G}_k$. The signal enhancement component, $\mathbf{G}_k$, is optimized by performing SVD to the transformed channel $\mathbf{H}_{B,k}^H\mathbf{F}_k$, yielding $\mathbf{H}_{B,k}^H\mathbf{F}_k = \tilde{\mathbf{V}}_k\tilde{\Lambda}_k\tilde{\mathbf{P}}_k^H$, where $\mathbf{H}_{B,k} \in \mathbb{C}^{N\times L}$ represents the channel matrix of the link from the BS to user k, where L denotes the number of antennas for each user, and $L \gg N$. The component $\mathbf{G}_k$ is typically selected as a scaled version of $\tilde{\mathbf{P}}_k$, such that $\mathbf{G}_k = (1/\sqrt{S})\tilde{\mathbf{P}}_k$, where S denotes the number of symbols to be transmitted to each user.

D. Adversarial Reflection Coefficient Optimization

This section examines an advanced adversarial jamming strategy where a malicious RIS manipulates downlink signals from the BS to disrupt legitimate communication, focusing on solving the associated nonconvex optimization problem. For analytical tractability, the system model described in Section IV-A is simplified to a single-user scenario. The malicious RIS is equipped with a microcontroller for channel estimation and reflection control, dynamically adjusting its reflection characteristics to maximize the disruption of the legitimate transmission. Each element of the malicious RIS can independently adjust its amplitude and phase. Specifically, the malicious RIS seeks to minimize the received signal power at the Rx by creating destructive interference between the direct and RIS-cascaded channels. This is achieved by jointly optimizing the reflection coefficient magnitudes (denoted by $\boldsymbol{\beta}$) and phase shifts (denoted by $\boldsymbol{\varphi}$) of its elements. The objective is equivalent to minimizing the SNR at the Rx, assuming constant noise power. The corresponding optimization problem is formulated as follows:

$$\min_{\boldsymbol{\beta},\boldsymbol{\varphi}} \ \left|\left(\mathbf{H}_{B,R}^H \Phi \mathbf{h}_R + \mathbf{h}_B\right)^H \mathbf{w}\right|^2 \tag{7a}$$

$$\text{s.t.} \ \ 0 \le \beta_m \le 1 \quad \forall m \in \{1,\ldots,M\} \tag{7b}$$

$$\varphi_m \in \Psi \quad \forall m \in \{1,\ldots,M\} \tag{7c}$$

where $\mathbf{h}_R \in \mathbb{C}^{M\times 1}$ denotes a channel vector from the RIS to Rx, $\mathbf{H}_{B,R} \in \mathbb{C}^{M\times N}$ is the channel matrix from the BS to the RIS, and $\mathbf{h}_B \in \mathbb{C}^{N\times 1}$ is the channel vector from the BS to Rx. $\mathbf{w} \in \mathbb{C}^{N\times 1}$ is the transmit beamforming vector at the BS, and $\Phi = \text{diag}(\beta_1 e^{j\varphi_1},\ldots,\beta_M e^{j\varphi_M}) \in \mathbb{C}^{M\times M}$ represents the diagonal reflection matrix of the RIS. $\boldsymbol{\beta} = [\beta_1,\ldots,\beta_M]^T$ is a vector of the reflection coefficient magnitudes, and $\boldsymbol{\varphi} = [\varphi_1,\ldots,\varphi_M]^T$ is a vector of phase shifts.

Due to the coupling between $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$, and the discrete constraint on $\boldsymbol{\varphi}$, Problem (7) is a mixed-integer nonlinear program (MINLP), which is generally NP-hard [60]. To tackle this, we propose a BCD framework that iteratively optimizes $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$, converging to a suboptimal solution that effectively minimizes the received signal power at Rx, thereby maximizing the jamming effectiveness.

1) Optimization of the Phase Shifts: For a fixed $\boldsymbol{\beta}$, the optimization of $\boldsymbol{\varphi}$ can be formulated as

$$\min_{\boldsymbol{\varphi}} \ \left|\left(\mathbf{H}_{B,R}^H \Phi \mathbf{h}_R + \mathbf{h}_B\right)^H \mathbf{w}\right|^2 \tag{8a}$$

$$\text{s.t.} \ \ \varphi_m \in \Psi \quad \forall m \in \{1,\ldots,M\}. \tag{8b}$$

To address the nonconvexity of (8), the discrete constraint on $\boldsymbol{\varphi}$ is first relaxed to allow continuous values, i.e., $0 \le \varphi_m \le 2\pi$. Define $x_m = e^{j\varphi_m}$ and $\mathbf{x} = [x_1,\ldots,x_M]^T$, where $|x_m| = 1 \ \ \forall m$. Substituting $\Phi = \text{diag}(\boldsymbol{\varphi}) = \text{diag}(\mathbf{x})\text{diag}(\boldsymbol{\beta})$ into the objective function (8a), and defining the $M \times 1$ vector $\mathbf{v}_1 = \text{diag}(\mathbf{h}_R)\mathbf{H}_{B,R}\mathbf{w} \odot \boldsymbol{\beta}$, and the scalar $v_2 = \mathbf{h}_B^H \mathbf{w}$, we can rewrite the objective as $|\mathbf{x}^H\mathbf{v}_1 + v_2|^2$, where $\odot$ denotes element-wise multiplication. Expanding the magnitude squared, we get the following optimization problem:

$$\min_{\mathbf{x}} \ \mathbf{x}^H\mathbf{v}_1\mathbf{v}_1^H\mathbf{x} + v_2^*\mathbf{x}^H\mathbf{v}_1 + v_2\mathbf{v}_1^H\mathbf{x} + |v_2|^2 \tag{9a}$$

$$\text{s.t.} \ \ |x_m| = 1 \quad \forall m \in \{1,\ldots,M\}. \tag{9b}$$

Problem (9) is still nonconvex. We employ an SDR technique combined with Gaussian randomization to tackle the unit modulus constraint.[4] We introduce an auxiliary vector $\mathbf{v}_4 = \begin{bmatrix}\mathbf{x}\\ 1\end{bmatrix} \in \mathbb{C}^{(M+1)\times 1}$ and auxiliary matrix $\mathbf{V}_3 = \begin{bmatrix}\mathbf{v}_1\mathbf{v}_1^H & \mathbf{v}_1 v_2^*\\ \mathbf{v}_1^H v_2 & 0\end{bmatrix} \in \mathbb{C}^{(M+1)\times(M+1)}$. We then introduce $\mathbf{V}_5 = \mathbf{v}_4\mathbf{v}_4^H = \begin{bmatrix}\mathbf{x}\mathbf{x}^H & \mathbf{x}\\ \mathbf{x}^H & 1\end{bmatrix} \in \mathbb{C}^{(M+1)\times(M+1)}$, which is a rank-1. The objective function (9a) can be rewritten as $\text{Tr}(\mathbf{V}_3\mathbf{V}_5) + |v_2|^2$. By relaxing the rank-1 constraint on $\mathbf{V}_5$, we obtain the semi-definite programming (SDP) formulation as

$$\min_{\mathbf{V}_5} \ \text{Tr}(\mathbf{V}_3\mathbf{V}_5) + |v_2|^2 \tag{10a}$$

$$\text{s.t.} \ \ V_{5,(m,m)} = 1 \quad \forall m \in \{1,\ldots,M+1\} \tag{10b}$$

$$\mathbf{V}_5 \succeq 0 \tag{10c}$$

where $V_{5,(m,m)}$ denotes the mth diagonal element of $\mathbf{V}_5$.

The SDP formulation is convex and can be solved using solvers, such as CVX [61]. However, the solution may not inherently satisfy the rank-one constraint. To derive a feasible solution, we employ Gaussian randomization [62]. The steps are as follows.

1) *EVD:* Let $\widetilde{\mathbf{V}}_5$ denotes the solution obtained from the SDP solver. Perform EVD: $\widetilde{\mathbf{V}}_5 = \mathbf{U}\mathbf{D}\mathbf{U}^H$, where $\mathbf{U} \in \mathbb{C}^{(M+1)\times(M+1)}$ is a unitary matrix, and $\mathbf{D} \in \mathbb{C}^{(M+1)\times(M+1)}$ is a diagonal matrix of eigenvalues.
2) *Randomization:* Generate a random vector $\mathbf{q} \in \mathbb{C}^{(M+1)\times 1}$ independently sampled from the complex normal distribution $\mathcal{CN}(0, \mathbf{I}_{M+1})$, where $\mathbf{I}_{M+1}$ is the identity matrix.
3) *Candidate Solution:* Construct a candidate vector: $\widehat{\mathbf{v}}_4 = \mathbf{U}\mathbf{D}^{1/2}\mathbf{q}$, where $\mathbf{D}^{1/2}$ is the element-wise square root of $\mathbf{D}$. Compute $\widehat{\mathbf{V}}_5 = \widehat{\mathbf{v}}_4\widehat{\mathbf{v}}_4^H$.
4) *Phase Extraction:* With $\widehat{\mathbf{v}}_4$, the phase shifts are determined as $\widehat{\boldsymbol{\varphi}} = -\arg([\widehat{\mathbf{v}}_4/\widehat{\mathbf{v}}_4(M+1)]_{(1,M)})$, where $[\widehat{\mathbf{v}}_4]_{(1,M)}$ represents the first M elements of $\widehat{\mathbf{v}}_4$, and $\widehat{\mathbf{v}}_4(M+1)$ denotes its $(M+1)$th element. The division is performed element-wise to ensure phase normalization. The transformation $[\widehat{\mathbf{v}}_4/\widehat{\mathbf{v}}_4(M+1)]_{(1,M)}$ maintains structural consistency with the formulation $\mathbf{v}_4 = \begin{bmatrix}\mathbf{x}\\ 1\end{bmatrix}$. The function $\arg(\mathbf{s})$ returns the phase angles of the complex-valued elements in $\mathbf{s}$. The negative sign ensures the RIS phases oppose the combined channel $\mathbf{v}_1$, thereby aligning the RIS-reflected signal $(\mathbf{x}^H\mathbf{v}_1)$ destructively with the direct channel (v_2), minimizing $|\mathbf{x}^H\mathbf{v}_1 + v_2|^2$.
5) *Quantization:* Map continuous phase $\widehat{\boldsymbol{\varphi}}$ to the nearest discrete value in the set Ψ: $\widehat{\boldsymbol{\varphi}^o} = \text{Quantize}(\widehat{\boldsymbol{\varphi}}, \Psi)$.
6) *Evaluation and Selection:* Repeat steps 2–5 for L randomization trials. Choose the $\widehat{\boldsymbol{\varphi}^o}$ that minimizes the objective function (8a). The selected $\widehat{\boldsymbol{\varphi}^o}$ will be $\boldsymbol{\varphi}$.

[4]While SDR provides a global lower bound for analyzing worst-case malicious RIS behavior in Problem (9), its high computational complexity limits scalability in large systems. As an alternative, Riemannian optimization methods, such as the RCG algorithm, operate directly on the complex circle manifold defined by unit-modulus constraints, offering feasible, memory-efficient, and scalable solutions. However, they converge to local optima, making them more suitable for practical scenarios than theoretical benchmarking. This tradeoff motivates our use of SDR for threat modeling, while highlighting Riemannian methods for scalable implementation.

2) Optimization of the Reflection Coefficient Magnitudes: With $\boldsymbol{\varphi}$ determined from the previous step, the optimization problem for $\boldsymbol{\beta}$ becomes

$$\min_{\boldsymbol{\beta}} \ |\left(\mathbf{H}_{B,R}^H \Phi \mathbf{h}_R + \mathbf{h}_B\right)^H \mathbf{w}|^2 \tag{11a}$$

$$\text{s.t.} \ \ 0 \le \beta_m \le 1 \quad \forall m \in \{1, \ldots, M\}. \tag{11b}$$

The objective function (11a) can be reformulated as $|\boldsymbol{\beta}^H \mathbf{c} + \mathbf{v}_2|^2$, where $\mathbf{c} = (\mathbf{H}_{B,R}^H \mathbf{x} \odot \mathbf{h}_R) \odot \mathbf{w}$ is a $M \times 1$ vector. Thus, the optimization problem becomes

$$\min_{\boldsymbol{\beta}} \ |\boldsymbol{\beta}^H \mathbf{c} + \mathbf{v}_2|^2 \tag{12a}$$

$$\text{s.t.} \ \ 0 \le \beta_m \le 1 \quad \forall m \in \{1, \ldots, M\}. \tag{12b}$$

The convexity of the Problem (12) ensures that a globally optimal solution for $\boldsymbol{\beta}$ can be efficiently obtained using standard solvers such as CVX.

Remarks: We propose a BCD framework that decouples the optimization of $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$, simplifying the optimization problem while ensuring convergence to a suboptimal solution. SDR, coupled with Gaussian randomization, provides an efficient and computationally tractable approach for handling discrete phase constraints. Despite the effectiveness of the proposed method in approximating the globally optimal solution, the performance is subject to problem-specific parameters such as the channel correlation structure, hardware impairments, and RIS element discretization levels. To further enhance performance, advanced techniques such as Riemannian manifold optimization, stochastic gradient-based learning, or DRL-based adaptive phase shift design can be explored.

E. Malicious Active RIS: Amplification for Enhanced Jamming

Active RISs address the inherent limitations of passive RISs, particularly the double-fading effect that attenuates signals in reflected paths, by integrating amplifiers into their elements. This enhancement enables active RISs to reflect and amplify incident signals, extending their operational range and influence on the wireless environment. However, this capability introduces tradeoffs, including increased power consumption and noise due to the active components [63]. Active RISs enhance jamming by amplifying and reflecting incident signals, extending their impact on the wireless environment. However, this comes with higher power consumption and amplified noise. Due to nonideal interelement isolation, active RISs are also prone to feedback-induced self-interference, which can distort destructive interference and reduce jamming precision [64]. Such interference may introduce detectable anomalies, potentially revealing malicious activity. As this work aims to establish a baseline system understanding, self-interference is not modeled in detail and is left for future investigation.

To explore the potential security threats posed by active RISs, we extend the single-user, single-antenna system model introduced in Section IV-D to include a malicious active RIS. Specifically, we examine how such a malicious entity can exploit its amplification capabilities to perform jamming through destructive beamforming, thereby degrading the SNR at the Rx. In this scenario, we consider a malicious active RIS equipped with a microcontroller capable of estimating CSI and dynamically adjusting the amplitude and phase of its reflected signals. Effective jamming requires the RIS to acquire accurate CSI. To achieve this, the malicious RIS initially operates under the guise of a legitimate RIS, collaborating with the BS during an initial phase. During this phase, the RIS provides feedback to the BS, enabling it to estimate the CSI of the BS-RIS ($\mathbf{H}_{B,R}$) and RIS-Rx ($\mathbf{h}_R$) links, as well as the direct BS-Rx channel ($\mathbf{h}_B$). The RIS, in turn, gains knowledge of these channel links and the transmit beamforming design $\mathbf{w}$.

The received signal at the legitimate Rx is

$$y = \underbrace{\left(\mathbf{H}_{B,R}^H \Phi \mathbf{h}_R + \mathbf{h}_B\right)^H \mathbf{w} s}_{\text{Desired signal}} + \underbrace{\mathbf{h}_R^H \Phi \mathbf{n}_R}_{\text{RIS noise}} + \underbrace{n_B}_{\text{Rx noise}} \tag{13}$$

where s is the transmitted symbol with unit power, $\mathbf{n}_R \sim \mathcal{CN}(0, \sigma_r^2 \mathbf{I}_M)$ is the additive noise at the malicious RIS, and $n_B \sim \mathcal{CN}(0, \sigma_B^2)$ is the additive noise at the Rx.

The malicious optimization problem can be formulated as

$$\max_{\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\varphi}} \ \frac{|\left(\mathbf{H}_{B,R}^H \Phi \mathbf{h}_R + \mathbf{h}_B\right)^H \mathbf{w}|^2}{||\mathbf{h}_R^H \Phi||^2 \sigma_r^2 + \sigma_B^2} \tag{14a}$$

$$\text{s.t.} \ \ \underbrace{||\mathbf{w}||^2}_{\text{BS Power}} + \underbrace{||\Phi^H \mathbf{H}_{B,R} \mathbf{w}||_2^2}_{\substack{\text{Amplified reflected} \\ \text{signal power}}} + \underbrace{\sigma_r^2 ||\Phi||_F^2}_{\text{RIS noise amplification}} \le P_T \tag{14b}$$

$$0 \le \beta_m \le \beta_m^{\max} \tag{14c}$$

$$\varphi_m \in [0, 2\pi] \quad \forall m \in \{1, \ldots, M\} \tag{14d}$$

where P_T is the total power budget, encompassing the transmit power of the BS and the power consumed by the active RIS [65], and $\beta_m^{\max}$ is the maximum allowable amplification factor for the mth RIS element. The norms $||.||_2^2$ and $||.||_F^2$ denote the squared Euclidean and Frobenius norms, respectively. For the diagonal matrix Φ with entries $\beta_m e^{j\varphi_m}$, $||\Phi||_F^2 = \sum_{m=1}^M \beta_m^2$.

The constraint (14b) captures the total power consumption, which includes the BS transmit power and the power consumed by the active RIS. The term $\|\Phi^H \mathbf{H}_{B,R} \mathbf{w}\|_2^2$ represents the power of the amplified reflected signal at the RIS, which depends solely on the incident signal, $\mathbf{H}_{B,R}\mathbf{w}$, and the RIS amplification configuration, Φ. The RIS-to-Rx channel, $\mathbf{h}_R$, does not appear in this expression, as it does not influence the RIS's power consumption; rather, it affects the received signal quality at the legitimate Rx, as reflected in the objective function (14a). This exclusion is consistent with the physical power model of active RIS systems. Problem (14) is challenging to solve due to its nonconvexity. The objective function (14a) is a fractional function with coupled variables, and the power constraint involves quadratic terms.

To address this, we consider two scenarios.

1) Scenario 1 (BS Aware of the Malicious RIS): When the BS is aware of the malicious RIS, it cannot rely on the channel estimates provided by the RIS. To design $\mathbf{w}$, the BS adopts a standard beamforming scheme, such as maximal ratio transmission (MRT), ZF, or signal-to-leakage-plus-noise ratio (SLNR), as summarized in Table IV [66]. After the BS

TABLE IV
BEAMFORMING SCHEMES AT THE BS AND THEIR OPERATIONAL CHARACTERISTICS

Beamforming Schemes	Operational Characteristics
MRT	• MRT maximizes the received signal strength at the intended user by aligning the transmitted signal with the user's channel. • Offers low computational complexity but may cause signal leakage to unintended users due to sidelobes, leading to interference.
ZF	• ZF minimizes inter-user interference by nulling the signal in the direction of other users, providing effective interference suppression. • This method can reduce diversity gain and has higher computational complexity, potentially impacting performance in environments with highly correlated channels.
SLNR	• SLNR optimizes the trade-off between maximizing the signal strength at the intended user and minimizing leakage to unintended users. • Balances interference suppression (similar to ZF) and signal enhancement (similar to MRT), providing robust performance in multi-user environments.

finalizes its beamforming strategy $\mathbf{w}$, the malicious RIS transitions into jamming mode. Utilizing its complete knowledge of the CSI (BS-RIS, RIS-Rx, and BS-Rx channels) and the BS-designed $\mathbf{w}$, the malicious RIS optimizes its $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ to minimize the Rx's SNR, which can be formulated as

$$\min_{\boldsymbol{\beta},\boldsymbol{\varphi}} \quad \frac{|\left(\mathbf{H}_{B,R}^H \Phi \mathbf{h}_R + \mathbf{h}_B\right)^H \mathbf{w}|^2}{||\mathbf{h}_R^H \Phi||^2 \sigma_r^2 + \sigma_B^2} \tag{15a}$$

$$\text{s.t.} \quad ||\Phi^H \mathbf{H}_{B,R} \mathbf{w}||_2^2 + \sigma_r^2 ||\Phi||_F^2 \leq P_{\text{RIS}} \tag{15b}$$

$$0 \leq \beta_m \leq \beta_m^{\max} \tag{15c}$$

$$\varphi_m \in [0, 2\pi] \quad \forall m \in \{1, \ldots, M\} \tag{15d}$$

where P_{RIS} is the power budget for the malicious active RIS.

Problem (15) can be solved using the BCD method described in Section IV-D, iteratively optimizing $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$.

2) Scenario 2 (BS Considers the Malicious RIS as Conventional RIS): In this more insidious scenario, the BS operates under the false premise that the RIS is a legitimate entity, purportedly collaborating to enhance the received SNR at the Rx. This deceptive coordination initiates a joint optimization of the BS beamforming vector $\mathbf{w}$ and the RIS reflection coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ to maximize the SNR received at the Rx. The corresponding maximization problem can be formulated as

$$\max_{\mathbf{w},\boldsymbol{\beta},\boldsymbol{\varphi}} \quad \frac{|\left(\mathbf{H}_{B,R}^H \Phi \mathbf{h}_R + \mathbf{h}_B\right)^H \mathbf{w}|^2}{||\mathbf{h}_R^H \Phi||^2 \sigma_r^2 + \sigma_B^2} \tag{16a}$$

$$\text{s.t.} \quad ||\mathbf{w}||^2 + ||\Phi^H \mathbf{H}_{B,R} \mathbf{w}||_2^2 + \sigma_r^2 ||\Phi||_F^2 \leq P_T \tag{16b}$$

$$0 \leq \beta_m \leq \beta_m^{\max} \tag{16c}$$

$$\varphi_m \in [0, 2\pi] \quad \forall m \in \{1, \ldots, M\}. \tag{16d}$$

Problem (16) is nonconvex due to the intricate coupling between $\mathbf{w}$, $\boldsymbol{\beta}$, and $\boldsymbol{\varphi}$. To tackle this complexity, an iterative approach can be used to find a locally optimal solution. This involves alternating between optimizing $\mathbf{w}$ for fixed $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$, and optimizing $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ for fixed $\mathbf{w}$.

Once the optimization parameters ($\mathbf{w}$, $\boldsymbol{\beta}$, and $\boldsymbol{\varphi}$) are determined based on this deceptive collaboration, the malicious RIS reveals its true nature and readjusts its $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ to minimize the SNR at Rx, using the formulation in (15).

Remarks: This two-stage attack, characterized by initial cooperation followed by malicious optimization, allows the active RIS to exploit the BS's trust and significantly degrade the Rx's communication quality. This highlights the potential security vulnerabilities introduced by active RISs and underscores the need for robust countermeasures.

Comparison of Active and Passive RIS Vulnerabilities: Though both active and passive RISs are vulnerable to adversarial manipulation, their distinct architectures lead to different threat profiles. Active RISs, with built-in amplifiers, offer enhanced jamming but introduce new attack surfaces, including amplifier nonlinearity and hardware tampering, which can disrupt beamforming or redirect signals [67]. Passive RISs, while more energy-efficient, are prone to phase-shift manipulation via control hijacking or electromagnetic interference, enabling stealthy attacks, such as eavesdropping. While passive RISs may allow prolonged undetected exploitation, active RISs are more likely to produce detectable anomalies (e.g., amplified noise). These differences highlight the need for architecture-specific defenses, e.g., robust hardware security for active RISs and enhanced anomaly detection for passive RISs.

V. PILOT CONTAMINATION ATTACK BY MALICIOUS RIS: NOVEL THREAT AND COUNTERMEASURES

Ensuring the confidentiality and integrity of wireless communication is paramount against increasingly sophisticated adversarial attacks. This section examines a novel and potent security threat posed by malicious RISs in TDD MIMO systems: *RIS-aided pilot contamination attacks* (PCAs). We analyze the unique challenges posed by these attacks, particularly the ineffectiveness of conventional PCA countermeasures. We then explore potential mitigation strategies to enhance PLS in the presence of malicious RISs.

A. Understanding RIS-Aided PCA

In TDD MIMO systems, the Tx typically acquires CSI by exploiting channel reciprocity. This is often done based on pilot signals transmitted from the Rx during an uplink training phase [68]. This process is vulnerable to PCAs, where an

TABLE V
CATEGORIES OF CONVENTIONAL PCA COUNTERMEASURES

Method	Cited Papers	Description	Key Points
Random Pilot Modulation	[69]	The pilot sequence $\mathbf{u} \in \mathbb{C}^{\tau_p \times 1}$ (τ_p is the sequence length) is divided into subsequences. Each subsequence is multiplied by a random symbol known only to the Rx.	• The Tx detects PCA by checking the cross-correlation between different channel estimates obtained from distinct pilot subsequences. • Detection is challenging because the Tx does not know the random symbols. • Securely sharing the random symbols between the Tx and Rx allows the Tx to estimate the CSIs of both the Rx and the Eve if PCA occurs.
Data/Noise Masking	[69]	In addition to $\mathbf{u}$, the Rx sends a random sequence $\mathbf{s} \in \mathbb{C}^{\tau_s \times 1}$, known only to the Rx.	• Two transmission methods for $\mathbf{s}$: superposition with $\mathbf{u}$ or separate transmission. • Eve's lack of knowledge of $\mathbf{s}$ ensures its transmitted sequence is linearly independent of the actual sequence sent by the Rx.
Dynamic Orthogonal Pilots	[70]	The Rx's pilot is randomly selected from a set of orthogonal pilot sequences.	• Eve is unaware of which pilot the Rx will send. • The Tx can detect PCA if multiple pilot signals are received simultaneously.
Statistical Feature Detection	[70]	PCA results in an increase in energy received by the Tx and a decrease in energy received by the Rx.	Energy-based detection is used to identify PCA.

Eve transmits the same pilot sequence as the Rx, interfering with the CSI estimation at the Tx. This corrupted CSI can then be exploited during the downlink data transmission phase, causing signal leakage toward Eve and degrading the performance of the Rx. Traditionally, countermeasures against PCA, as summarized in Table V, involve techniques, such as:

1) *Random Pilot Modulation:* introducing randomness into pilot signals to differentiate them from malicious replicas;
2) *Data/Noise Masking:* adding artificial noise or data to mask the pilot sequence and make it harder to replicate;
3) *Dynamic Orthogonal Pilots:* employing dynamically changing orthogonal pilot sequences to prevent Eve from consistently mimicking the legitimate pilot; and
4) *Statistical Feature Detection:* exploiting statistical differences in received signals (e.g., power levels and timing variations) to detect the presence of a PCA.

However, the emergence of RIS technology introduces a new dimension to this threat and renders many traditional countermeasures ineffective [71]. A malicious RIS, controlled by Eve, can passively reflect the pilot signal transmitted by the Rx back toward the Tx. This reflected pilot signal is a time-synchronized and phase-coherent replica of the legitimate pilot, making it virtually indistinguishable from the intended signal. Because the RIS simply reflects the legitimate pilot sequence without modification or the need for complex signal processing, techniques such as random pilot modulation, data/noise masking, and dynamic orthogonal pilots become obsolete. This necessitates the development of novel countermeasures specifically designed to mitigate RIS-aided PCA.

B. Challenges and Potential Countermeasures

RIS-aided PCA presents two key challenges.

1) Detection of RIS-Aided PCA: Identifying the presence of an RIS-aided PCA is challenging because the reflected pilot signal is, by design, nearly identical to the legitimate one. Traditional statistical feature-based methods, which rely on analyzing discrepancies in received signal characteristics, may struggle to differentiate between the two. More advanced techniques are needed.

1) *Exploiting Spatial Correlation:* Analyzing the spatial correlation of the received signal at the Tx to differentiate between a direct path signal and a reflected signal. This leverages the fact that the direct and reflected paths will likely have distinct spatial signatures (e.g., different angles of arrival, angular spreads, or path diversities).
2) *Leveraging Temporal Consistency:* This involves exploiting the temporal correlation properties of the channel. While legitimate channels typically exhibit slow and predictable variations, RIS-induced reflections may change more rapidly or display irregular variations if the RIS actively manipulates the reflection coefficients.
3) *Advanced Detection Techniques:*
   a) *Beyond Block-Based Detection:* Current research on PCA detection predominantly relies on binary hypothesis testing within each channel coherence block. However, this approach may be suboptimal when Eve launches cumulative PCA over multiple blocks. More sophisticated detection strategies are as follows.
      i) *Sequential Detection:* Continuously monitoring the received signal over multiple blocks to accumulate evidence and achieve more accurate detection.
      ii) *ML-Based Detection:* Training machine learning (ML) models to recognize patterns and features associated with RIS-aided PCA, enabling more intelligent and adaptive detection mechanisms [72]. These models could learn subtle differences in signal statistics or

temporal behavior that are difficult to capture with traditional methods.

b) *Exploiting RIS Behavior:* Investigating the properties of the reflected signal to gather information about Eve's location or channel state characteristics. In some scenarios, it might be possible to turn the RIS against the attacker by exploiting this information.

c) *Higher-Order Statistics:* Examining higher-order statistical moments (e.g., skewness and kurtosis) of the received signal beyond the mean and variance can aid in detecting subtle anomalies introduced by RIS-induced perturbations in the channel.

2) Secure Transmission After PCA Detection: Even if RIS-aided PCA is detected, securing data transmission in the downlink remains challenging. Conventional beamforming techniques, such as MRT, rely on accurate CSI. However, in the presence of PCA, the estimated CSI is corrupted, rendering these techniques suboptimal [73]. Advanced channel estimation and beamforming schemes are therefore crucial for maintaining secure data transmission.

1) *Robust Beamforming:* Designing beamforming vectors that are less sensitive to CSI errors and minimize signal power toward the estimated direction of the malicious RIS. This approach reduces information leakage by preventing the RIS from effectively reflecting the signal toward Eve. Potential strategies are as follows.

a) *RIS-Aware Beamforming:* Developing algorithms that explicitly account for the potential presence of a malicious RIS during the beamforming design process.

b) *ML for Adaptive Beamforming:* Leveraging advanced ML models, such as reinforcement learning or deep neural networks, to dynamically adapt beamforming strategies, identifying RIS-induced perturbations on the channel and iteratively optimizing beamforming weights to nullify the influence of the malicious RIS.

2) *Artificial Noise Injection:* Intentionally adding artificial noise to the transmitted signal can degrade Eve's channel estimate while minimally impacting legitimate Rx. This technique can further reduce information leakage by making it harder for Eve to exploit the contaminated CSI.

The emergence of RIS-aided PCA underscores the evolving nature of wireless security threats. Addressing this challenge requires a paradigm shift in how we approach CSI acquisition, beamforming design, and attack detection. By actively pursuing the research directions outlined in this section, we can develop secure and resilient wireless communication systems that can effectively counter the threats posed by intelligent adversaries equipped with RIS technologies. This includes developing novel signaling schemes, robust beamforming strategies, and advanced detection mechanisms that can adapt to the dynamic and adversarial nature of future wireless networks.

VI. Eavesdropping Attacks by Malicious RIS

While jamming degrades the SINR at the Rx to disrupt legitimate communications, eavesdropping aims to intercept the information. Traditional RIS-aided security systems focus on enhancing legitimate communications by mitigating interference and improving QoS for legitimate users. However, the adaptability of RIS technologies also introduces the potential for adversarial use, wherein a malicious RIS can be deployed to aid eavesdropping. It poses a significant threat to wireless security by dynamically reconfiguring its reflection coefficients to manipulate signal propagation, thereby creating favorable conditions for interception. Unlike legitimate RIS, designed to optimize communication for legitimate users, a malicious RIS can introduce deliberate propagation anomalies to redirect and improve signals toward an Eve [74]. This strategic manipulation enables the interception of signals that would otherwise be inaccessible due to obstacles, distance, or attenuation, significantly reducing the secrecy rate of the system and undermining the confidentiality of transmitted data.

This section explores novel strategies and mechanisms for eavesdropping enabled by malicious RIS, emphasizing their implications for 6G wireless networks.

A. Cooperative Eavesdropping With Multiple Malicious RISs

A single malicious RIS might have limited coverage for effective eavesdropping, especially if Tx and Rx are geographically separated or if the direct path between Tx and Eve is weak. To overcome this limitation, multiple malicious RISs can cooperate to create a distributed beamforming network, effectively extending the eavesdropping range and enhancing the intercepted signal strength at Eve [75].

1) *Strategic Placement:* Malicious RISs are strategically positioned to intercept and redirect the signal. This might involve placing RISs near the Tx, the Rx, or along a carefully chosen path to create a multihop reflection-based forwarding trajectory toward the Eve.

2) *Coordinated Beamforming and Amplification:* Each RIS in the network adjusts its reflection coefficients (phase shifts and the amplification factors for the active RIS) in a coordinated manner to perform distributed beamforming. This focuses the intercepted signal energy toward the next RIS in the chain or, ultimately, toward Eve. Active RISs can further amplify the signal to compensate for path loss and improve the SNR at the Eve.

This cooperative eavesdropping strategy necessitates careful coordination and synchronization among the malicious RISs, posing several operational challenges.

1) *Time Synchronization:* To coherently combine the reflected signals, RISs must operate in a time-synchronized manner. Achieving precise synchronization across multiple distributed RISs can be technically demanding.

2) *Channel Estimation:* For effective beamforming, Eve needs to estimate all the involved channels. This can be challenging due to the complexity of cascaded channels

involving multiple RISs. Eve might attempt to estimate these channels using techniques, such as analyzing pilot signals (if available) or exploiting potential uplink transmissions, assuming some degree of channel reciprocity.

3) *Coordinated Control:* Efficiently coordinating the reflection coefficients of multiple RISs requires a control mechanism, either via a centralized controller with communication links to each RIS or a distributed algorithm where each RIS makes decisions based on local information and limited communication with neighboring RISs. Detecting such coordinated activity serves as a potential countermeasure for legitimate users. This might involve analyzing the statistical properties of received signals for anomalies or employing specialized probing techniques to reveal the presence of multiple cooperating RISs.

B. Exploiting RIS for Side-Channel Information Leakage

Even without directly intercepting the transmitted data, a malicious RIS can be leveraged to gather sensitive side-channel information, potentially compromising the privacy and security of communication systems. By acting as a controllable perturbation node within the wireless channel, a malicious RIS enables an Eve to infer critical information through advanced analysis of signal perturbations and variations [77]. Key mechanisms facilitating this leakage are as follows.

1) *Channel Probing:* By strategically manipulating its reflection coefficients, a malicious RIS can inject controlled perturbations into the legitimate channel. Analyzing the impact of these perturbations on the received signal (either at a compromised legitimate user or at Eve) allows Eve to extract information about the legitimate users. This can include location and mobility, antenna configuration, modulation schemes, and user activity.
2) *Timing Analysis:* Precise measurement of the time-of-flight of signals reflected by the RIS enables triangulation and tracking of communicating devices, revealing their locations and movement patterns. Accuracy can be enhanced with Angle-of-Arrival (AoA) estimation.
3) *Correlation Attacks:* Eve can correlate subtle changes in the RIS-reflected signal with external events (e.g., user activity) or environmental factors (e.g., temperature changes). Over time, these correlations can reveal information about the environment and user behavior.

Protecting against such side-channel attacks requires a multifaceted approach, incorporating robust signaling schemes, randomized and adaptive transmission patterns, real-time RIS detection and tracking, and dynamic CSI monitoring for anomaly detection. Further research is critical to design targeted countermeasures tailored to the unique vulnerabilities introduced by RIS-aided side-channel attacks.

C. RIS-Aided Attacks on Physical Layer Secret Key Generation

PLS often relies on the inherent channel reciprocity between the Tx and Rx for secret key generation. This reciprocal nature of wireless channels implies that Tx and Rx observe identical channel characteristics when exchanging signals. However, an RIS, controlled by Eve, can disrupt this process, undermining the reliability of key generation mechanisms [76].

1) *Asymmetric Channel Manipulation:* A malicious RIS enables an attack by creating asymmetric channel conditions. By manipulating its reflection coefficients, it ensures that the channel Tx observes when transmitting to Rx differs from the channel Rx observes when transmitting to Tx. This disruption of channel reciprocity causes mismatched secret keys. Since Eve controls a malicious RIS, it can exploit its configuration to reconstruct the original channel or predict the manipulated channels as seen by either the Tx or Rx. This allows Eve to derive the secret key, breaking secure communication and potentially decrypting subsequent transmissions [77].
2) *Countermeasures and Future Research Directions:* New RIS-aware key agreement protocols are essential, incorporating mechanisms to actively probe the channel for inconsistencies, leveraging higher-order statistical moments (beyond mean and variance) of the channel measurements, and integrating authentication methods to verify the integrity of channel measurements [78]. Further research is needed to analyze and quantify the impact of RIS manipulation on secrecy rates. ML techniques offer promising tools for reliably detecting and mitigating RIS-induced channel distortions. Additionally, thoroughly investigating the limits of secure key generation with malicious RISs, including theoretical bounds on achievable key generation rates under various attack scenarios, is critical. Addressing these challenges will enable the development of robust PLS key generation techniques to counter the growing threat of malicious RISs.

D. Enhancing Eve's SNR via Malicious RIS Reflection

To analyze the innovative eavesdropping strategies of a malicious RIS, we consider an MISO system, as described in Section IV-D. In this scenario, a BS with N antennas transmits confidential information to a single-antenna Rx. A malicious RIS with M reflecting elements, controlled by a single-antenna Eve, can manipulate its reflection coefficients to enhance the received signal strength at Eve and intercept the BS-Rx communications. These strategies operate under the common assumption that the malicious RIS knows the BS-Eve, BS-RIS, and RIS-Eve channels.

The received SNR at Eve can be expressed as

$$\gamma_e = \frac{|\left(\mathbf{H}_{B,R}^H \Phi \mathbf{h}_{R,E} + \mathbf{h}_{B,E}\right)^H \mathbf{w}|^2}{\sigma_e^2} \tag{17}$$

where $\mathbf{h}_{B,E} \in \mathbb{C}^{N\times 1}$ is the channel vector from the BS to Eve, $\mathbf{h}_{R,E} \in \mathbb{C}^{M\times 1}$ is the channel vector from the RIS to Eve, and σ_e^2 is the variance of AWGN at Eve.

The RIS aims to maximize the received SNR at Eve, γ_e, by optimizing its reflection coefficient matrix, Φ. We explore several optimization strategies for the RIS.

1) Phase-Only Optimization (Passive RIS): The RIS adjusts only the phase shifts, setting $\beta_m = 1$ for all m (i.e., $\boldsymbol{\beta} = \mathbf{1}$). The optimization problem becomes

$$\max_{\boldsymbol{\varphi}} \ \gamma_e \tag{18a}$$
$$\text{s.t.} \ \ \varphi_m \in [0, 2\pi] \quad \forall m \in \{1, \ldots, M\}. \tag{18b}$$

This problem is nonconvex due to the presence of phase-dependent variables. It can be tackled using advanced optimization techniques, such as SDR with Gaussian randomization, gradient-based approaches, or heuristic algorithms (e.g., particle swarm optimization or genetic algorithms) [79].

2) Joint Amplitude and Phase Optimization (Passive RIS): Here, the RIS has more control and can optimize both the amplitude and phase of its reflection coefficients. For a passive RIS, the amplitude is constrained between 0 and 1 ($0 \leq \beta_m \leq 1$ for all m, i.e., $\mathbf{0} \leq \boldsymbol{\beta} \leq \mathbf{1}$). The optimization problem is

$$\max_{\boldsymbol{\beta}, \boldsymbol{\varphi}} \ \gamma_e \tag{19a}$$
$$\text{s.t.} \ \ 0 \leq \beta_m \leq 1 \tag{19b}$$
$$\varphi_m \in [0, 2\pi] \quad \forall m \in \{1, \ldots, M\}. \tag{19c}$$

Problem (19) is challenging due to the coupling between the amplitude and phase variables, making it highly nonconvex. Advanced optimization techniques, such as BCD, alternating optimization, or SDR, can derive suboptimal solutions. These methods iteratively decouple and optimize the amplitude and phase variables to address the computational complexity.

3) Active RIS Optimization: An active RIS can amplify the reflected signals, potentially enhancing eavesdropping capabilities even further [77]. The optimization problem, considering the total power budget and the noise introduced by its active elements, is formulated as

$$\max_{\boldsymbol{\beta}, \boldsymbol{\varphi}} \ \frac{|\left(\mathbf{H}_{B,R}^H \Phi \mathbf{h}_{R,E} + \mathbf{h}_{B,E}\right)^H \mathbf{w}|^2}{\|\mathbf{h}_{RE}^H \Phi\|^2 \sigma_r^2 + \sigma_e^2} \tag{20a}$$
$$\text{s.t.} \ \ ||\Phi^H \mathbf{H}_{B,R} \mathbf{w}||_2^2 + \sigma_r^2 ||\Phi||_F^2 \leq P_{\text{RIS}} \tag{20b}$$
$$0 \leq \beta_m \leq \beta_m^{\max} \tag{20c}$$
$$\varphi_m \in [0, 2\pi] \quad \forall m \in \{1, \ldots, M\}. \tag{20d}$$

This problem is similar to optimization problems encountered in jamming scenarios. It can be solved using advanced iterative optimization methods tailored to handle the coupled amplitude, phase, and power constraints.

Remarks: Imperfect CSI is a major challenge in these scenarios, as malicious RIS may not know precisely all channels involved. Robust optimization techniques, such as worst-case or probabilistic approaches, can address the uncertainties associated with imperfect CSI. Furthermore, the optimization process must adapt to dynamic channel variations and the behavior of the legitimate Tx and Rx. Finally, the possibility of collaboration between the malicious RIS and Eve could further enhance channel estimation and signal processing capabilities, creating a more challenging threat landscape.

E. Nonreciprocal Eavesdropping With ND-RIS

This strategy leverages ND-RIS to facilitate eavesdropping. Unlike conventional RISs with diagonal reflection matrices that treat each element as an independent scatterer, ND-RIS employs a nondiagonal matrix, enabling interelement coupling for sophisticated signal manipulation [79]. This breaks the inherent reciprocity of the wireless channel, creating asymmetry between uplink and downlink channels.

1) Channel Reciprocity Disruption: During the uplink channel estimation phase, the ND-RIS remains passive, allowing the BS to estimate the uplink channel, $\mathbf{h}_{\text{UL}}$, based on pilot signals transmitted from the Rx. This process typically relies on the assumption of channel reciprocity. During downlink transmission, the ND-RIS activates a nondiagonal reflection matrix Φ. The reflected signal from each element becomes a weighted linear combination of the incident signals on all elements rather than a simple phase-shifted and scaled version of the signal incident on that specific element. This intricate interplay of signals breaks the reciprocity between the uplink and downlink channels. The BS, unaware of ND-RIS manipulation, designs the downlink precoding vector $\mathbf{w}$ based on the estimated uplink channel $\mathbf{h}_{\text{UL}}$, assuming reciprocity. However, ND-RIS alters the actual downlink channel $\mathbf{h}_C$, leading to suboptimal precoding. The nonreciprocity is mathematically expressed as $\mathbf{h}_C \neq \mathbf{h}_{\text{UL}}^H$.

2) Secrecy Rate Optimization: The malicious ND-RIS aims to exploit this discrepancy by optimizing its nondiagonal reflection matrix Φ to minimize the achievable secrecy rate, defined as the difference between the achievable rate at the Rx and the Eve. The optimization problem can be formulated as

$$\min_{\Phi} \ C_s = \log_2\left(1 + \frac{|\mathbf{h}_C^H \mathbf{w}|^2}{\sigma_B^2}\right) - \log_2\left(1 + \frac{\mathbf{h}_{C,E}^H \mathbf{w}|^2}{\sigma_e^2}\right) \tag{21a}$$
$$\text{s.t.} \ \ 0 \leq \beta_m \leq 1 \tag{21b}$$
$$\varphi_m \in [0, 2\pi] \quad \forall m \in \{1, \ldots, M\} \tag{21c}$$

where $\mathbf{h}_C = \mathbf{H}_{B,R}^H \Phi \mathbf{h}_R + \mathbf{h}_B \in \mathbb{C}^{N \times 1}$, and $\mathbf{h}_{C,E} = \mathbf{H}_{B,R}^H \Phi \mathbf{h}_{R,E} + \mathbf{h}_{B,E} \in \mathbb{C}^{N \times 1}$ are the combined downlink channels to the legitimate Rx and Eve, respectively.

Problem (21) is inherently nonconvex due to the nondiagonal structure of Φ and the secrecy rate objective. Solving it requires advanced optimization techniques.

1) *Iterative Element-Wise Optimization:* Optimize each element of Φ iteratively while keeping other elements fixed. This can be combined with gradient-based or heuristic methods. However, global optimality is not guaranteed.
2) *Manifold Optimization:* Exploits the phase constraints of Φ by searching on the complex unit circle manifold, effectively handling the nonconvex nature.
3) *Alternating Optimization:* Decomposes the original nonconvex problem into lower-dimensional subproblems, optimizing them iteratively in an alternating fashion.

Remarks: The solution approach depends on the structural design and control capabilities of ND-RIS, the imposed constraints, and the available computational resources. An active ND-RIS can also be considered. However, appropriate power

constraints must be incorporated, with their specific formulation depending on the underlying active ND-RIS architecture.

VII. Case Study: Null-Steering at Rx and Constructive Interference at Eve

We present a case study examining a malicious RIS employing a sophisticated strategy that extends beyond simply maximizing the signal received by an Eve, as discussed in Section VI-D. By creating a signal null in the direction of the legitimate Rx while steering the reflected signal constructively toward Eve, the RIS can simultaneously degrade Rx's signal quality and enhance Eve's channel. This dual-purpose approach integrates null-steering and beamforming techniques for adversarial intent, inducing deep fading at Rx while improving Eve's reception. To implement this attack, we formulate a single optimization problem that balances null-steering at Rx and signal enhancement for Eve, controlled by a tradeoff weighting factor λ (where $0 \leq \lambda \leq 1$)

$$\min_{\boldsymbol{\beta},\boldsymbol{\varphi}} \ \lambda\left|\left(\mathbf{H}_{B,R}^{H}\Phi\mathbf{h}_{R}+\mathbf{h}_{B}\right)^{H}\mathbf{w}\right|^{2} -(1-\lambda)\left|\left(\mathbf{H}_{B,R}^{H}\Phi\mathbf{h}_{R,E}+\mathbf{h}_{B,E}\right)^{H}\mathbf{w}\right|^{2} \tag{22a}$$

$$\text{s.t.} \ \ 0 \leq \beta_m \leq 1 \tag{22b}$$

$$\varphi_m \in [0, 2\pi] \quad \forall m \in \{1, \ldots, M\}. \tag{22c}$$

The weighting factor λ determines the relative importance of the two objectives and can be adjusted to achieve the desired tradeoff based on the specific scenario. This strategy resembles null-space projection in multiuser MIMO systems where the RIS manipulates the reflected signal to align the combined signal at Rx with a subspace that attenuates the direct signal, effectively creating a null. In basic secrecy rate terms, the key concern is whether Eve's received power exceeds Rx's to achieve a positive secrecy rate for the adversary. However, this analysis delves deeper by incorporating specific constraints to uncover the full adversarial potential of a malicious RIS. By leveraging the tradeoff parameter λ, this study reveals how the RIS can be tuned to prioritize different objectives, such as minimizing Rx's power while ensuring Eve meets a power threshold or maximizing Eve's power while capping Rx's power. This detailed approach provides actionable insights into the RIS's dual capability to disrupt legitimate communication and enhance eavesdropping, which is vital for system designers seeking to mitigate such threats. We assume perfect channel knowledge at the RIS; in practice, channel estimation errors would reduce the attack's efficacy. For an active RIS, additional constraints, such as power budget limitations and active noise amplification, must be incorporated into the optimization problem.

To solve the nonconvex Problem (22), iterative optimization methods can be employed. One approach is *alternating optimization*, where $\boldsymbol{\beta}$ is first fixed while optimizing $\boldsymbol{\varphi}$ by computing the gradient of the objective function. The phase shifts are adjusted along the negative gradient for null-steering and the positive gradient for Eve enhancement, scaled according to λ. Once $\boldsymbol{\varphi}$ is updated, $\boldsymbol{\beta}$ is optimized while keeping $\boldsymbol{\varphi}$ fixed, and the process is repeated iteratively until convergence. Another method involves *gradient-based optimization*, where the gradient of the entire objective function is utilized to iteratively update both variables. Additionally, *heuristic or convex relaxation techniques* can be applied to achieve computationally efficient yet near-optimal solutions [80]. In this work, we employ gradient-based optimization using MATLAB's `fmincon` with the interior-point algorithm, which efficiently solves constrained nonlinear problems via iterative convex approximations. To enhance robustness, we use randomized initialization and average results over multiple channel realizations. Although global optimality is not guaranteed due to nonconvex RIS constraints, the method consistently yields reliable local solutions. We also acknowledge that heuristic approaches, such as SDR, remain effective and widely used alternatives for large-scale RIS optimization.

To assess the effectiveness of this strategy, two specific constraints are introduced to reflect practical adversarial scenarios.

Constraint 1: Minimize the received power at Rx while ensuring Eve receives a minimum power level of $E_{\min}$

$$\min_{\boldsymbol{\beta},\boldsymbol{\varphi}} \ \left|\left(\mathbf{H}_{B,R}^{H}\Phi\mathbf{h}_{R}+\mathbf{h}_{B}\right)^{H}\mathbf{w}\right|^{2} \tag{23a}$$

$$\text{s.t.} \ \ \left|\left(\mathbf{H}_{B,R}^{H}\Phi\mathbf{h}_{R,E}+\mathbf{h}_{B,E}\right)^{H}\mathbf{w}\right|^{2} \geq E_{\min} \tag{23b}$$

$$0 \leq \beta_m \leq 1, \quad \varphi_m \in [0, 2\pi] \quad \forall m \in \{1, \ldots, M\}. \tag{23c}$$

Constraint 2: Maximize the received power at Eve while keeping the power at Rx below a maximum threshold of $R_{\max}$

$$\max_{\boldsymbol{\beta},\boldsymbol{\varphi}} \ \left|\left(\mathbf{H}_{B,R}^{H}\Phi\mathbf{h}_{R,E}+\mathbf{h}_{B,E}\right)^{H}\mathbf{w}\right|^{2} \tag{24a}$$

$$\text{s.t.} \ \ \left|\left(\mathbf{H}_{B,R}^{H}\Phi\mathbf{h}_{R}+\mathbf{h}_{B}\right)^{H}\mathbf{w}\right|^{2} \leq R_{\max} \tag{24b}$$

$$0 \leq \beta_m \leq 1, \quad \varphi_m \in [0, 2\pi] \quad \forall m \in \{1, \ldots, M\}. \tag{24c}$$

These constraints enable exploration of the malicious RIS's operational limits, identifying specific λ values where the system optimally balances disruption of Rx and enhancement of Eve's reception. The simulation results derived under these constraints offer key performance insights into the RIS's behavior and its potential impact on adversarial wireless systems.

A. Simulation Setup

We model an MISO system where the BS, equipped with N antennas, transmits to a single-antenna Rx in the presence of a single-antenna Eve. A malicious passive RIS with M reflecting elements is strategically positioned to manipulate the signal propagation environment. We assume perfect channel knowledge at the RIS to establish a theoretical upper bound on the adversarial impact of malicious RIS configurations. In practice, channel estimation errors would reduce the attack's efficacy, and future work could explore this aspect. The simulation employs a quasi-static Rayleigh fading model, where channels are assumed constant within each transmission frame but vary independently across frames. Results are

TABLE VI
SIMULATION PARAMETERS

Parameter	Description
N	6
M	40
λ	Varied from 0 to 1 in steps of 0.01
Transmit Power (P_t)	40 W
Path Loss Exponent (γ)	2.5
Reference Path Loss (C_0)	-20 dB (at 1 m)
RIS Type	Passive
Channel Model	Rayleigh fading

averaged over 100 independent channel realizations to capture the statistical behavior of time-varying channels. For scenarios with sample-by-sample channel variation, auto-regressive (AR) models could be adopted to account for fading within a packet [81]. The legitimate Rx and Eve are equidistant from the BS, with distances $d_B = d_{B,E} = 80$ m. The RIS is positioned midway between the BS and Rx/Eve, such that the distances are $d_{B,R} = d_R = d_{R,E} = 25$ m. This configuration ensures the RIS is closer to both BS and Rx/Eve than the direct BS-Rx/Eve distance, enabling the reflected path to interact constructively or destructively with the direct link. Wireless channels follow Rayleigh fading, with amplitudes scaled by path loss based on the given distances. The BS employs MRT beamforming to optimize the signal for Rx, assuming no knowledge of the RIS's presence. The beamforming vector is $\mathbf{w} = (\mathbf{h}_B/\|\mathbf{h}_B\|)\sqrt{P_t}$. For each λ, φ_m, and β_m are optimized using MATLAB's `fmincon` interior-point algorithm for the objective function (22a). The simulation parameters are summarized in Table VI.

The two constraints are incorporated as follows.

1) *Constraint 1:* Minimize Rx power subject to Eve power $\geq E_{\min} = 0.0030$ W. This models an attack that disrupts Rx while maintaining a minimum eavesdropping capability, which is practical for degrading Rx performance without fully sacrificing Eve's interception ability.
2) *Constraint 2:* Maximize Eve power subject to Rx power $\leq R_{\max} = 0.00065$ W. This prioritizes eavesdropping effectiveness while ensuring Rx's signal remains sufficiently degraded, reflecting a strategy where interception is the primary goal with controlled Rx disruption.

These constrained optimization problems are solved to identify critical λ values, offering insights into the tradeoffs and operational boundaries of the malicious RIS.

B. Discussion

Fig. 3 illustrates the variation of received powers at Rx and Eve as λ ranges from 0 to 1. As λ increases, both Rx and Eve powers decrease. Specifically, Rx power decreases monotonically from its maximum at $\lambda = 0$ to its minimum at $\lambda = 1$, reflecting the RIS's increasing emphasis on null-steering at Rx. Similarly, Eve's power decreases with rising λ, as the RIS's efforts to disrupt Rx inadvertently reduce Eve's signal strength due to shared channel characteristics between Rx and Eve. This behavior underscores the RIS's need to balance its dual adversarial objectives, requiring careful tuning of λ to meet specific attack goals. Vertical lines in Fig. 3 indicate the λ values where predefined constraints are satisfied:

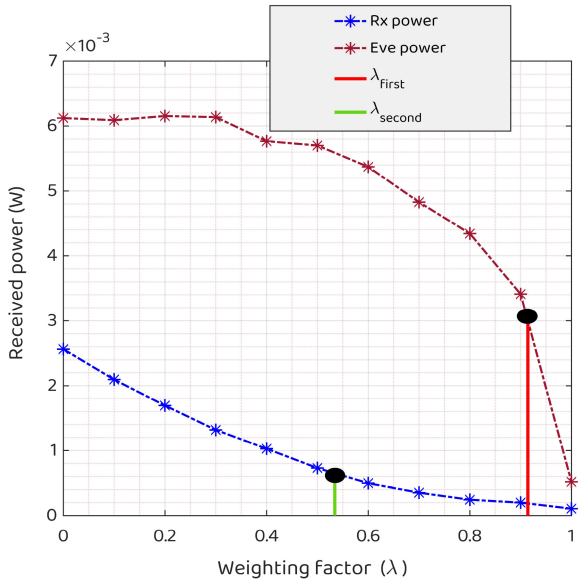


Fig. 3. Received power at Rx and Eve. depending on a weighting factor λ.

at $\lambda = 0.53$, Constraint 2 is met with Rx power = 0.00065 W and Eve power = 0.0057 W; at $\lambda = 0.91$, Constraint 1 is met with Rx power = 0.00032 W and Eve power = 0.0030 W. The tradeoff curve in Fig. 4 is obtained by plotting the average received power at Rx against that at Eve for each value of λ. For each λ, φ_m and β_m are jointly optimized by solving the optimization problem (22). The curve is constructed by connecting the resulting power pairs, with Rx power plotted on the x-axis and Eve power on the y-axis. This process maps the relationship between Rx and Eve powers across the full range of λ, revealing the tradeoffs inherent in the RIS's adversarial capabilities. Two specific points on the curve are highlighted, as they satisfy the predefined constraints.

1) *Red Marker (Constraint 1):* At $\lambda = 0.91$, the RIS reduces the Rx power to 0.00032 W while maintaining Eve's power above 0.0030 W, demonstrating its ability to induce deep fading at the legitimate Rx.
2) *Green Marker (Constraint 2):* At $\lambda = 0.53$, the RIS boosts Eve's power to 0.0057 W while keeping the legitimate Rx power below 0.00065 W, highlighting its potential to enable targeted eavesdropping.

These constraint-satisfying points are determined through linear interpolation between adjacent λ values, ensuring the power thresholds specified by Constraints 1 and 2 are accurately met. The resulting tradeoff curve in Fig. 4 illustrates the inverse relationship between Rx and Eve powers, highlighting the dual adversarial role of the RIS. The identified λ values (i.e., $\lambda = 0.53$ and $\lambda = 0.91$) serve as practical benchmarks for system designers, illustrating how the malicious RIS can be tuned to prioritize different attack objectives. These results underscore the critical need for robust defenses against RIS-enabled threats in future 6G networks.

VIII. CONCLUSION AND FUTURE WORK

This article has presented a pioneering investigation into the emerging security vulnerabilities introduced by malicious

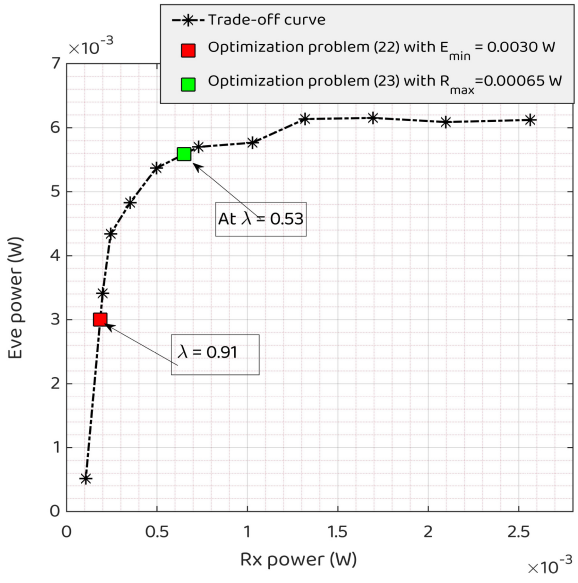


Fig. 4. Tradeoff curve between received power at Rx and Eve.

RISs in 6G networks. Departing from the conventional focus on the performance benefits of RIS, we have systematically explored its adversarial potential, identifying and analyzing key threat vectors, such as sophisticated jamming, eavesdropping, and PCAs. Through rigorous analytical modeling and scenario-driven evaluations, we demonstrated the feasibility and disruptive impact of these RIS-enabled attacks, emphasizing how adversaries could exploit advanced optimization strategies and system-level design weaknesses to compromise network integrity. These findings underscore the dual-use nature of RIS technology and the pressing need for robust security-aware frameworks.

Future research must prioritize the development of proactive and scalable defense mechanisms, including resilient detection frameworks and secure deployment protocols tailored to RIS-specific vulnerabilities. Concrete directions include ML-driven anomaly detection, hardware-level authentication schemes, cross-layer security integration, and quantum-resistant RIS control mechanisms. While the detailed design and comprehensive performance evaluation of such countermeasures are beyond the scope of this initial threat-focused study, they represent critical future research directions motivated by our findings. As 6G networks evolve, ensuring their resilience will depend on our ability to anticipate, model, and mitigate the emerging threats introduced by malicious RIS through collaborative standardization, intelligent adaptation, and cross-disciplinary security solutions.

REFERENCES

[1] X. Chen, J. Tan, L. Kang, F. Tang, M. Zhao, and N. Kato, "Frequency selective surface toward 6G communication systems: A contemporary survey," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 3, pp. 1635–1675, 3rd Quart., 2024.

[2] M.A. Siddiqi, H. Yu, and J. Joung, "5G ultra-reliable low-latency communication implementation challenges and operational issues with IoT devices," *Electronics*, vol. 8, no. 9, p. 981, Sep. 2019.

[3] K. Yan, W. Ma, Q. Yang, S. Sun, and W. Wang, "Info-chain: Reputation-based blockchain for secure information sharing in 6G intelligent transportation systems," *IEEE Internet Things J.*, vol. 11, no. 5, pp. 9198–9212, Mar. 2024.

[4] C. De Alwis, P. Porambage, K. Dev, T. R. Gadekallu, and M. Liyanage, "A survey on network slicing security: Attacks, challenges, solutions and research directions," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 1, pp. 534–570, Feb. 2024.

[5] M. Wei, H. Zhao, V. Galdi, L. Li, and T. J. Cui, "Metasurface-enabled smart wireless attacks at the physical layer," *Nat. Electron.*, vol. 6, no. 8, pp. 610–618, Aug. 2023.

[6] Z. Lin et al., "Refracting RIS-aided hybrid satellite-terrestrial relay networks: Joint beamforming design and optimization," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 4, pp. 3717–3724, Aug. 2022.

[7] W. Khalid, A.-A. A. Boulogeorgos, T. Van Chien, J. Lee, H. Lee, and H. Yu, "Optimal operation of active RIS-aided wireless powered communications in IoT networks," *IEEE Internet Things J.*, vol. 12, no. 1, pp. 390–401, Jan. 2025.

[8] S. Noh, K. Seo, Y. Sung, D. J. Love, J. Lee, and H. Yu, "Joint direct and indirect channel estimation for RIS-assisted millimeter-wave systems based on array signal processing," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 8378–8391, Nov. 2023.

[9] S. Noh, J. Lee, G. Lee, K. Seo, Y. Sung, and H. Yu, "Channel estimation techniques for RIS-assisted communication: Millimeter-wave and sub-THz systems," *IEEE Veh. Technol. Mag.*, vol. 17, no. 2, pp. 64–73, Jun. 2022.

[10] M. Ahmed et al., "A Survey on STAR-RIS: Use cases, recent advances, and future research challenges," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14689–14711, Aug. 2023.

[11] W. Khalid, H. Yu, J. Cho, Z. Kaleem, and S. Ahmad, "Rate-energy tradeoff analysis in RIS-SWIPT systems with hardware impairments and phase-based amplitude response," *IEEE Access*, vol. 10, pp. 31821–31835, 2022.

[12] M. Shahjalal et al., "Enabling technologies for AI-empowered 6G massive radio access networks," *ICT Exp.*, vol. 9, no. 3, pp. 341–355, Jun. 2023.

[13] W. Khalid, Z. Kaleem, R. Ullah, T. V. Chien, S. Noh, and H. Yu, "Simultaneous transmitting and reflecting-reconfigurable intelligent surface in 6G: Design guidelines and future perspectives," *IEEE Netw.*, vol. 37, no. 5, pp. 173–181, Dec. 2022.

[14] M. Misbah, Z. Kaleem, W. Khalid, C. Yuen, and A. Jamalipour, "Phase and 3-D placement optimization for rate enhancement in RIS-assisted UAV networks," *IEEE Wireless Commun. Lett.*, vol. 12, no. 7, pp. 1135–1138, Jul. 2023.

[15] W. Khalid, H. Yu, D.-T. Do, Z. Kaleem, and S. Noh, "RIS-aided physical layer security with full-duplex jamming in underlay D2D networks," *IEEE Access*, vol. 9, pp. 99667–99679, 2021.

[16] L. Zhi et al., "Self-powered absorptive reconfigurable intelligent surfaces for securing satellite-terrestrial integrated networks," *China Commun.*, vol. 21, no. 9, pp. 276–291, Sep. 2024.

[17] J. Bae, W. Khalid, A. Lee, H. Lee, S. Noh, and H. Yu, "Overview of RIS-enabled secure transmission in 6G wireless networks," *Digit. Commun. Netw.*, vol. 10, no. 6, pp. 1553–1565, Dec. 2024.

[18] H. Yu, T. Kim, and H. Jafarkhani, "Wireless secure communication with beamforming and jamming in time-varying wiretap channels," *IEEE Trans. Inf. Forensics Security*, vol. 13, pp. 2087–2100, 2018.

[19] S. M. S. Shahriyer, A. S. M. Badrudduza, S. Shabab, M. K. Kundu, and H. Yu, "Opportunistic relay in multicast channels with generalized shadowed fading effects: A physical layer security perspective," *IEEE Access*, vol. 9, pp. 155726–155739, 2021.

[20] A. S. M. Badrudduza et al., "Security at the physical layer over GG fading and mEGG turbulence induced RF-UOWC mixed system," *IEEE Access*, vol. 9, pp. 18123–18136, 2021.

[21] H. Yu and I.-G. Lee, "Physical layer security based on NOMA and AJ for MISOSE channels with an untrusted relay," *Future Gener. Comput. Syst.*, vol. 102, pp. 611–618, Jan. 2020.

[22] S. Kim and H. Yu, "Energy-efficient HARQ-IR for massive MIMO systems," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3892–3901, Sep. 2018.

[23] E. Illi et al., "Physical layer security for authentication, confidentiality, and malicious node detection: A paradigm shift in securing IoT networks," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 1, pp. 347–388, Feb. 2024.

[24] W. Khalid, H. Yu, and S. Noh, "Residual energy analysis in cognitive radios with energy harvesting UAV under reliability and secrecy constraints," *Sensors*, vol. 20, no. 10, pp. 2998–3017, May 2020.

[25] H. Yu and T. Kim "Training and data structures for AN-aided secure communication," *IEEE Syst. J.*, vol. 13, no. 3, pp. 2869–2872, Sep. 2019.

[26] H. Yu and J. Joung, "Design of the power and dimension of artificial noise for secure communication systems," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4001–4010, Jun. 2021.
[27] W. Khalid and H. Yu, "Security improvement with QoS provisioning using service priority and power allocation for NOMA-IoT networks," *IEEE Access*, vol. 9, pp. 9937–9948, 2021.
[28] G. C. Alexandropoulos et al., "RIS-enabled smart wireless environments: Deployment scenarios, network architecture, bandwidth and area of influence," *EURASIP J. Wireless Commun. Netw.*, vol. 2023, no. 1, p. 103, Oct. 2023.
[29] W. U. Khan et al., "Opportunities for physical layer security in UAV communication enhanced with intelligent reflective surfaces," *IEEE Wireless Commun.*, vol. 29, no. 6, pp. 22–28, Dec. 2022.
[30] M. H. Khoshafa et al., "RIS-assisted physical Layer security in emerging RF and optical wireless communication systems: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, early access, Oct. 28, 2024, doi: 10.1109/COMST.2024.3487112.
[31] M. M. Rahman, A. S. M. Badrudduza, N. A. Sarker, M. Ibrahim, I. S. Ansari, and H. Yu, "RIS-aided mixed RF-FSO wireless networks: Secrecy performance analysis with simultaneous eavesdropping," *IEEE Access*, vol. 11, pp. 126507–126523, 2023.
[32] A. B. Sarawar, A. S. M. Badrudduza, M. Ibrahim, I. S. Ansari, and H. Yu, "Secrecy Performance analysis of integrated RF-UOWC IoT networks enabled by UAV and underwater-RIS," *IEEE Internet Things J.*, vol. 12, no. 3, pp. 2592–2608, Feb. 2025.
[33] Y. Sun et al., "Multi-functional RIS-assisted semantic anti-jamming communication and computing in integrated aerial-ground networks," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 12, pp. 3597–3617, Dec. 2024.
[34] W. Khalid, M. A. U. Rehman, T. Van Chien, Z. Kaleem, H. Lee, and H. Yu, "Reconfigurable intelligent surface for physical layer security in 6G-IoT: Designs, issues, and advances," *IEEE Internet Things J.*, vol. 11, no. 2, pp. 3599–3613, Jan. 2024.
[35] X. Luo and H. Zhu, "IRS-based TDD reciprocity breaking for pilot decontamination in massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 10, no. 1, pp. 102–106, Jan. 2021.
[36] L. Dai, H. Huang, C. Zhang, and K. Qiu, "Silent flickering RIS aided covert attacks via intermittent cooperative jamming," *IEEE Wireless Commun. Lett.*, vol. 12, no. 6, pp. 1027–1031, Jun. 2023.
[37] Y. Wang, H. Lu, D. Zhao, Y. Deng, and A. Nallanathan, "Wireless communication in the presence of illegal reconfigurable intelligent surface: Signal leakage and interference attack," *IEEE Wireless Commun.*, vol. 29, no. 3, pp. 131–138, Jun. 2022.
[38] B. Lyu, D. T. Hoang, S. Gong, D. Niyato, and D. I. Kim, "IRS-based wireless jamming attacks: When jammers can attack without power," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1663–1667, Oct. 2020.
[39] D. Gürgünoğlu, E. Björnson, and G. Fodor, "Combating inter-operator pilot contamination in reconfigurable intelligent surfaces assisted multi-operator networks," *IEEE Trans. Commun.*, vol. 72, no. 9, pp. 5884–5895, Sep. 2024.
[40] C.-K. Wen, L.-S. Tsai, A. Shojaeifard, P.-K. Liao, K.-K. Wong, and C.-B. Chae, "Shaping a smarter electromagnetic landscape: IAB, NCR, and RIS in 5G standard and future 6G," *IEEE Commun. Stand. Mag.*, vol. 8, no. 1, pp. 72–78, Mar. 2024.
[41] S. Rivetti, Ö. T. Demir, E. Björnson, and M. Skoglund, "Malicious reconfigurable intelligent surfaces: How impactful can destructive beamforming be?" *IEEE Wireless Commun. Lett.*, vol. 13, no. 7, pp. 1918–1922, Jul. 2024.
[42] W. Khalid, H. Yu, R. Ali, and R. Ullah, "Advanced physical-layer technologies for beyond 5G wireless communication networks," *Sensors*, vol. 21, no. 9, p. 3197, May 2020.
[43] Z. Wei, W. Hu, J. Zhang, W. Guo, and J. A. McCann, "Explainable adversarial learning framework on physical layer key generation combating malicious reconfigurable intelligent surface," *IEEE Trans. Wireless Commun.*, vol. 24, no. 4, pp. 3529–3545, Apr. 2025.
[44] C. Zou et al., "Multi-layer RIS-assisted anti-jamming communications: A hierarchical game learning approach," *IEEE Commun. Lett.*, vol. 27, no. 11, pp. 2998–3002, Nov. 2023.
[45] X. Gan et al., "Bayesian learning for double-RIS aided ISAC systems with superimposed pilots and data," *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 5, pp. 766–781, Jul. 2024.
[46] H. Alakoca et al., "Metasurface manipulation attacks: Potential security threats of RIS-aided 6G communications," *IEEE Commun. Mag.*, vol. 61, no. 1, pp. 24–30, Jan. 2023.
[47] B. D. Son et al., "Adversarial attacks and defenses in 6G network-assisted IoT systems," *IEEE Internet Things J.*, vol. 11, no. 11, pp. 19168–19187, Jun. 2024.
[48] H. Wang, Z. Han, and A. L. Swindlehurst, "Channel reciprocity attacks using intelligent surfaces with nondiagonal phase shifts," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 1469–1485, 2024.
[49] G. Li et al., "RIS-Jamming: Breaking key consistency in channel reciprocity-based key generation," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 5090–5105, 2024.
[50] M. Abdollahi, K. Malekinasab, W. Tu, and M. Bag-Mohammadi, "Physical-layer jammer detection in multihop IoT networks," *IEEE Internet Things J.*, vol. 10, no. 23, pp. 20574–20585, Dec. 2023.
[51] G. Marti, T. Kölle, and C. Studer, "Mitigating smart jammers in multi-user MIMO," *IEEE Trans. Signal Process.*, vol. 71, pp. 756–771, 2023.
[52] X. Wang, C. Hua, and Y. Qiu, "Event-triggered model-free adaptive control for non-linear multiagent systems under jamming attacks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 14458–14466, Oct. 2024.
[53] Q. Li et al., "Reconfigurable intelligent surfaces relying on non-diagonal phase shift matrices," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6367–6383, Jun. 2022.
[54] W. Khalid and H. Yu, "Optimal sensing performance for cooperative and non-cooperative cognitive radio networks," *Int. J. Distrib. Sensor Netw.*, vol. 13, no. 11, pp. 1–9, Nov. 2017.
[55] H. Huang, Y. Zhang, H. Zhang, Y. Cai, A. L. Swindlehurst, and Z. Han, "Disco Intelligent Reflecting Surfaces: Active channel aging for fully-passive jamming attack," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 806–819, Jan. 2024.
[56] I. Singh, P. J. Smith, and P. A. Dmochowski, "Imperfect CSI analysis in RIS-aided wireless systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 19475–19488, Dec. 2024.
[57] H. Huang, Y. Zhang, H. Zhang, C. Zhang, and Z. Han, "Illegal intelligent reflecting surface based active channel aging: When jammer can attack without power and CSI," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 11018–11022, Aug. 2023.
[58] Q. Li, M. El-Hajjar, I. Hemadeh, A. Shojaeifard, and L. Hanzo, "Coordinated reconfigurable intelligent surfaces: Non-diagonal group-connected design," *IEEE Trans. Veh. Technol.*, vol. 73, no. 7, pp. 10811–10816, Jul. 2024.
[59] A. S. de Sena, J. Kibiłda, N. H. Mahmood, A. Gomes, and M. Latva-Aho, "Malicious RIS versus massive MIMO: Securing multiple access against RIS-based jamming attacks," *IEEE Wireless Commun. Lett.*, vol. 13, no. 4, pp. 989–993, Apr. 2024.
[60] N. Khan, A. Ahmad, A. Wakeel, Z. Kaleem, B. Rashid, and W. Khalid, "Efficient UAVs deployment and resource allocation in UAV-relay assisted public safety networks for video transmission," *IEEE Access*, vol. 12, pp. 4561–4574, 2024.
[61] Z. Peng, R. Weng, C. Pan, G. Zhou, M. D. Renzo, and A. L. Swindlehurst, "Robust transmission design for RIS-assisted secure multiuser communication systems in the presence of hardware impairments," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 7506–7521, Nov. 2023.
[62] Y. Wen, G. Chen, S. Fang, Z. Chu, P. Xiao, and R. Tafazolli, "STAR-RIS-Assisted-Full-Duplex jamming design for secure wireless communications system," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 4331–4343, 2024.
[63] Z. Lin et al., "Pain without gain: Destructive beamforming from a malicious RIS perspective in IoT networks," *IEEE Internet Things J.*, vol. 11, no. 5, pp. 7619–7629, Mar. 2024.
[64] Z. Zhang et al., "Active RIS vs. passive RIS: Which will prevail in 6G?" *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1707–1725, Mar. 2023.
[65] W. Khalid, H. Yu, and A.-A. A. Boulogeorgos, "Outage analysis for active reconfigurable intelligent surface-enhanced wireless powered communication networks," in *Proc. IEEE Int. Conf. Artif. Intel. Inf. Commun. (ICAIIC)*, Osaka, Japan, Feb. 2024, pp. 473–475.
[66] W. Guo, Y. Lu, H. Du, B. Ai, D. Niyato, and Z. Ding, "Hybrid MRT and ZF learning for energy-efficient transmission in multi-RIS-assisted networks," *IEEE Trans. Veh. Technol.*, vol. 73, no. 8, pp. 12247–12251, Aug. 2024.
[67] M. Ahmed et al., "Active reconfigurable intelligent surfaces: Expanding the frontiers of wireless communication-A survey," *IEEE Commun. Surveys Tuts.*, vol. 27, no. 2, pp. 839–869, Apr. 2025.
[68] K.-W. Huang and H.-M. Wang, "Intelligent reflecting surface aided pilot contamination attack and its countermeasure," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 345–359, Jan. 2021.
[69] W. Zhang, H. Lin, and R. Zhang, "Detection of pilot contamination attack based on uncoordinated frequency shifts," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2658–2670, Jun. 2018.

[70] W. Wang, N. Cheng, K. C. Teh, X. Lin, W. Zhuang, and X. Shen, "On countermeasures of pilot spoofing attack in massive MIMO systems: A double channel training based approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 6697–6708, Jul. 2019.
[71] H.-M. Wang, K.-W. Huang, and T. A. Tsiftsis, "Multiple antennas secure transmission under pilot spoofing and jamming attack," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 4, pp. 860–876, Apr. 2018.
[72] N. Gao, Z. Qin, and X. Jing, "Pilot contamination attack detection and defense strategy in wireless communications," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 938–942, Jun. 2019.
[73] H. Fang, H. Hu, Y. Zhang, L. Yang, and H. Zhu, "Joint design of pilot power and phase shifts in RIS-aided cell-free massive MIMO URLLC systems," *IEEE Internet Things J.*, vol. 11, no. 22, pp. 37399–37402, Nov. 2024.
[74] G. C. Alexandropoulos, K. D. Katsanos, M. Wen, and D. B. Da Costa, "Counteracting eavesdropper attacks through reconfigurable intelligent surfaces: A new threat model and secrecy rate optimization," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1285–1302, 2023.
[75] M. R. A. Ruku, M. Ibrahim, A. S. M. Badrudduza, I. S. Ansari, W. Khalid, and H. Yu, "Effects of co-channel interference on RIS empowered wireless networks amid multiple eavesdropping attempts," *ICT Exp.*, vol. 10, no. 3, pp. 491–497, Jun. 2024.
[76] L. Hu, G. Li, A. Hu, and D. W. K. Ng, "Exploiting malicious RIS for secret key acquisition in physical-layer key generation," *IEEE Wireless Commun. Lett.*, vol. 13, no. 2, pp. 417–421, Feb. 2024.
[77] T. Van Chien, L. T. Tu, W. Khalid, H. Yu, S. Chatzinotas, and M. Di Renzo, "RIS-assisted wireless communications: long-term versus short-term phase shift designs," *IEEE Trans. Commun.*, vol. 72, no. 2, pp. 1175–1190, Oct. 2023.
[78] S.-E. Jeon, S.-J. Lee, Y.-R. Lee, H. Yu, and I.-G. Lee, "Machine learning-based cooperative clustering for detecting and mitigating jamming attacks in beyond 5G networks," *Inf. Syst. Front.*, 2024, to be published, doi: 10.1007/s10796-024-10534-6.
[79] Z. Kaleem, W. Khalid, A. Ahmad, H. Yu, A. M. Almasoud, and C. Yuen, "Reinforcement learning for energy-efficient user association in UAV-assisted cellular networks," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 60, no. 2, pp. 2474–2481, Apr. 2024.
[80] Z. Kaleem, W. Khalid, A. Muqaibel, A. A. Nasir, C. Yuen, and G. K. Karagiannidis, "Learning-aided UAV 3D placement and power allocation for sum-capacity enhancement under varying altitudes," *IEEE Commun. Lett.*, vol. 26, no. 7, pp. 1633–1637, Jul. 2022.
[81] H. Yu and J. Joung, "Optimization of frame structure and fronthaul compression for uplink C-RAN under time-varying channels," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1278–1292, Feb. 2021.

Waqas Khalid (Member, IEEE) received the B.S. degree in electronics engineering from the GIK Institute of Engineering Sciences and Technology, Topi, Pakistan, in 2011, the M.S. degree in information and communication engineering from Inha University, Incheon, South Korea, in 2016, and the Ph.D. degree from Yeungnam University, Gyeongsan, South Korea, in 2019.

He is currently a Research Professor with the Institute of Industrial Technology, Korea University, Seoul, South Korea. His areas of interest include physical-layer modeling, signal processing, and emerging technologies for 5G networks, including reconfigurable intelligent surfaces, physical-layer security, NOMA, cognitive radio, UAVs, and IoTs.

Trinh Van Chien received the B.S. degree in electronics and telecommunications from Hanoi University of Science and Technology (HUST), Hanoi, Vietnam, in 2012, the M.S. degree in electrical and computer engineering from Sungkyunkwan University, Seoul, South Korea, in 2014, and the Ph.D. degree in communication systems from Linköping University, Linköping, Sweden, in 2020.

He was a Research Associate with the University of Luxembourg, Esch-sur-Alzette, Luxembourg. He is currently with the School of Information and Communication Technology, HUST. His research interests include convex optimization problems and machine learning applications for wireless communications.

Dr. Chien also received the Award of Scientific Excellence in the First Year of the 5G Wireless Project funded by European Union Horizon's 2020. He was an Exemplary Reviewer of IEEE WIRELESS COMMUNICATIONS LETTERS in 2016, 2017, and 2021, and IEEE TRANSACTIONS ON COMMUNICATIONS in 2022.

Wali Ullah Khan (Member, IEEE) received the master's degree in electrical engineering from COMSATS University Islamabad, Islamabad, Pakistan, in 2017, and the Ph.D. degree in information and communication engineering from Shandong University, Qingdao, China, in 2020.

He is currently with the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg City, Luxembourg. He has authored/co-authored more than 150 publications, including international journals, peer-reviewed conferences, and book chapters. His research interests include convex/nonconvex optimizations, terrestrial and nonterrestrial wireless networks, reflecting intelligent surfaces, ambient backscatter communications, the Internet of Things, holographic MIMO, joint sensing and communication, physical layer security, and applications of machine learning.

Zeeshan Kaleem (Senior Member, IEEE) received the B.S. degree in electrical engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2007, the M.S. dgeree in electrical, electronics, controls and instrumentation Engineering from Hanyang University, Seoul, South Korea, in 2010, and the Ph.D. degree in electronics engineering from INHA University, Incheon, South Korea, in 2016.

He is currently serving as an Assistant Professor with the Computer Engineering Department, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia. Before joining KFUPM, he was an Associate Professor with COMSATS University Islamabad, Wah Campus, Pakistan. He authored/co-authored more than 100 technical papers and patents.

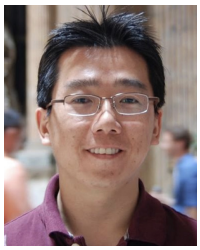
Dr. Kaleem was the recipient of the National Research Productivity Award from the Pakistan Council of Science and Technology, and the Higher Education Commission Best Innovator Award in 2017. He is an Editor of IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and serving as the Track Chair for IEEE VTC-Spring 2025. He served as an Editor for several journals, including *IEEE Communications Magazine*, IEEE WIRELESS COMMUNICATIONS, and IEEE ACCESS and served as the TPC Member for several flagship conferences.

Yousaf Bin Zikria (Senior Member, IEEE) received the Ph.D. degree from the Department of Information and Communication Engineering, Yeungnam University, Gyeongsan, South Korea, in 2016.

From 2018 to 2022, he was an Assistant Professor with the Department of Information and Communication Engineering, College of Engineering, Yeungnam University. He is currently a Program Coordinator for the IT and the Cyber Security programs with TIIS Sydney, NSW, Australia. With over 100 published papers and more than 6000 Google citations, he is listed among the World'S Top 2% of Researchers from 2021 to 2024 published in Elsevier by Stanford University.

Taejoon Kim (Member, IEEE) received the B.S. degree in electronics engineering from Yonsei University, Seoul, Republic of Korea, in 2003, and the Ph.D. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 2011.

He is currently a Full Professor with the School of Information and Communication Engineering, Chungbuk National University, Cheongju, Republic of Korea. From 2003 to 2005, he was a Researcher with LG Electronics, Seoul, Republic of Korea. In 2011, he was a Postdoctoral Researcher with KAIST. From 2011 to 2013, he was a Senior Researcher with ETRI, Daejeon, Republic of Korea. His research areas include the optimization of wireless networks with machine learning, reinforcement learning for automated reasoning, and generative artificial intelligence.



Heejung Yu (Senior Member, IEEE) received the B.S. degree in radio science and engineering from Korea University, Seoul, South Korea, in 1999, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2001 and 2011, respectively.

From 2001 to 2012, he was with the Electronics and Telecommunications Research Institute, Daejeon, and from 2012 to 2019, he was with Yeungman University, Gyeongsan, South Korea. He is currently a Professor with the Department of Electronics and Information Engineering, Korea University, Sejong, South Korea. His research interests include statistical signal processing and communication theory.