Risk-Aware Adaptive Cyber Deception Guided by Large Language Models

David Lopes Antunes[1][0009−0000−5274−1570], Pavlos Cheimonidis[2][0009−0002−4253−1232], Eleftherios Batzolis[2][0000−0002−5411−2716], Kyriakos Ovaliadis[2][0000−0003−2454−5630],, Salvador Llopis Sanchez[1][0000−0003−3828−4136], and Konstantinos Rantos[2][0000−0003−2453−3904]

[1] Universidad Politecnica de Valencia, Spain
{daan3,salllosa}@upv.edu.es
[2] Department of Informatics, Democritus University of Thrace, Greece
{pcheimon,egbatzo,ovaliad,krantos}@cs.duth.gr

Abstract. Modern cyber threats demand defense strategies that are adaptive, risk-aware, and capable of misdirecting adversaries in real time. Traditional static deception systems lack the flexibility to respond to evolving attack patterns and changing mission priorities. To address this, we introduce a framework for risk-aware adaptive cyber deception assisted by Large Language Models. The architecture integrates dynamic risk assessment, AI-assisted deception strategy generation, and modular deployment mechanisms. At its core, the Decision and Policy Engine uses an LLM-driven agent to interpret MITRE CAPEC-aligned threat intelligence and generate semantically rich deception recommendations. These are then translated into executable deception playbooks by the Dynamic Cyber Deception module, which manages tactic selection and deployment. The framework includes a feedback loop where telemetry and mission impact assessments inform ongoing refinement of deception strategies, enabling mission-aware adaptation over time. This work lays a foundation for the next generation of intelligent cyber defense systems that combine structured risk models with language-model reasoning to support resilient, adaptive, and context-driven deception capabilities.

Keywords: AI-driven Cyber Deception · Dynamic Risk Assessment · Common Attack Pattern Enumeration and Classification· Deception Scheme · Adaptive Deception Playbooks

1 Introduction

Modern cyber-attacks are increasingly sophisticated, making it harder for organizations to defend against them using traditional security mechanisms. Static defences, like firewalls and antivirus software, are no longer enough to keep pace with the rapidly changing threat landscape. Attackers constantly adapt their techniques, leaving defenders struggling to identify and mitigate new threats quickly. To counter this, security systems must become more dynamic, intelligent, and proactive.

Dynamic Risk Assessment (DRA) provides a flexible and responsive approach to assessing cybersecurity risks by continuously updating threat probabilities and asset vulnerabilities based on real-time data [5]. Rather than depending solely on pre-defined rules or historical attack patterns, DRA techniques leverage probabilistic models like Bayesian Networks to assess the current threat landscape for the target environment. These models integrate internal data, such as network topologies and deployed assets, with external threat intelligence, like vulnerability databases and exploit prediction scores. In particular, public resources such as the Common Vulnerabilities and Exposures (CVE) catalog, the National Vulnerability Database (NVD), the Common Weakness Enumeration (CWE), and the Common Attack Pattern Enumeration and Classification (CAPEC) offer structured information about known vulnerabilities, software weaknesses, and attacker behaviours. By combining these sources, dynamic risk models deliver up-to-date assessments of which systems are most at risk at any given time, enabling more informed and adaptive decision-making.

Cyber deception, on the other hand, offers another proactive defence strategy by deliberately creating and managing deceptive elements such as honeypots, honeytokens, decoy applications, and fake data [12]. The goal of deception is to mislead and confuse attackers, delay their progress, and gather intelligence about their methods. Effective cyber deception increases the attacker's workload and uncertainty, making successful exploitation more difficult and costly. However, in many current systems, deception assets are deployed statically, without adapting to the evolving risk environment, and thus reducing their potential impact.

Cyber deception has the potential to contribute to the full spectrum of cyberspace operations (from purely Defence to Offense), especially in proactive and reactive defence against sophisticated cyber threats. In this paper, we utilise aspects of the taxonomy developed by Lopez *et al.* [4], and more specifically the defined tactics and techniques, and provide an enhanced set of categories for the deception mechanisms.

Their used AI-driven framework spans all defence phases: prevention, detection, reaction, and forensics, unlike previous limited-scope approaches. AI, especially Machine Learning and Deep Learning, is underutilized in Cyber Deception despite its potential to enhance deception adaptability and precision from Cybersecurity to Cyber Defence contexts. There are still key challenges to take into account, such as the absence of standardized evaluation metrics, limited use of multi-technique deception, insufficient attention to stealth and offensive tactics and the configuration of the underlying network infrastructure [11]. In this paper we adopt a more granular categorisation of cyber deception mechanisms than the one proposed in [4], comprising the following categories:

- Detect: this type of deception is designed to spot attacker early in the reconnaissance phase. It is used as early warning systems.
- Mislead (or misdirect): the goal is to trick attackers into going the wrong way. It is a manoeuvre in cyberspace for diverting the attacker.
- Engage (or interact): the goal is to prolong the interaction with the adversary to collect valuable intelligence.

– Respond: This type of deception aims to project power in cyberspace (offensive operations).

While dynamic risk assessment and cyber deception have each been studied extensively, existing approaches typically treat them as separate efforts. Risk models are often used to prioritize patches or firewall rules, while deception strategies are deployed independently, based on general assumptions about attacker behavior. There is a critical need for a unified methodology that uses dynamic, real-time risk insights to formulate and drive an organization's cyber deception strategy.

In this paper, we propose a novel methodology that connects dynamic risk assessment with the strategic planning and implementation of cyber deception. Our framework analyzes the results of continuous risk assessment to identify high-risk scenarios — situations where specific vulnerabilities, threat actors, or attack paths present the greatest danger. Based on these high-risk scenarios, we guide the selection and deployment of appropriate deception mechanisms, tailoring the deceptive environment to the current threat landscape. This creates a more responsive, intelligent defence posture that evolves as risks change over time.

The main contributions of this paper are:

– A novel adaptive cyber defence framework that unifies Dynamic Risk Assessment (DRA) with context-aware cyber deception to enable dynamic, mission-aligned defensive actions.
– A methodology for the formulation of cyber deception strategies by leveraging real-time high-risk scenario analysis combined with Large Language Models (LLMs) and specialised prompt engineering.
– A modular architecture for the selection, orchestration, and adaptive deployment of deception mechanisms, integrating semantic deception playbooks and dynamic feedback loops.
– An operational pipeline that links asset-level risk probabilities, CAPEC-based threat patterns, and deception scheme generation into a closed-loop, mission-driven cyber defense system.

The rest of this paper is structured as follows. Section 2 reviews related work. Section 3 presents our proposed methodology. Sections 4 and 5 conclude with a discussion of future work.

2 Related work

Recent research has increasingly focused on proactive cyber defence approaches, particularly dynamic risk assessment and deception and the combination of these two has started attracting the research community.

Sengupta *et al.* [22] survey Moving Target Defence (MTD) techniques for network security, categorizing them based on what, when, and how movement

occurs. They highlight the role of artificial intelligence and SDN/NFV technologies in enabling dynamic defences but do not directly integrate dynamic risk models with deception planning.

Li *et al.* [16] propose an optimal defensive deception framework for container-based cloud environments using Deep Reinforcement Learning (DRL). Their work focuses on adaptively deploying decoys based on system dynamics, but emphasizes placement optimization rather than a systematic methodology linking dynamic risk insights to broader deception strategy formulation.

De Faveri *et al.* [8] develop a multi-paradigm modeling approach to incorporate deception tactics into software design processes. Although this supports early-stage security engineering, it lacks a connection to dynamic operational risk updates or runtime deception adaptation.

Al-Shaer *et al.* [11] introduce the notion of autonomous cyber deception using dynamic decision-making frameworks, deep learning, and HoneyThings. Their work envisions automated deception, but focuses primarily on generating deceptive artifacts rather than structuring deception strategies based on evolving risk profiles.

Wang *et al.* [23] design a proactive deception decision-making model using Bayesian attack graphs and Stackelberg games to optimize honeypoint placement. While they integrate dynamic attack path analysis with deception deployment, their approach concentrates mainly on optimizing decoy allocation rather than developing deception strategies at the scenario level.

Huang and Zhu [10] present a multi-stage dynamic Bayesian game model to counter Advanced Persistent Threats (APTs) in cyber-physical systems. Their framework captures stealthy attacker-defender interactions, incorporating proactive defence and deception, but does not propose an explicit methodology for translating dynamic risk evaluation into deception planning.

In contrast to these prior efforts, our work introduces a novel methodology that systematically links dynamic risk assessment outputs to the formulation and implementation of an organization's cyber deception strategy. By identifying high-risk scenarios in real time, our framework leverages a Large Language Model (LLM) to assist in generating tailored deception strategies that dynamically align deception mechanisms with the current threat landscape. This integration of dynamic risk-driven reasoning, LLM-assisted strategy generation, and scenario-specific deception deployment represents a significant advancement over prior approaches that treat risk modeling and deception independently or statically.

3 Proposed Framework

Figure 1 depicts the proposed framework following a modular architecture where the various modules must interact in a sequential order. The framework operates as a closed-loop system linking real-time risk assessment to deception planning and deployment. The Information Collection and DRA modules identify high-risk assets and associated CAPEC attack patterns. The Decision and Policy Engine (DPE) interprets these CAPECs using an LLM-based specialist agent,

producing tailored deception strategies structured as Deception Strategy Reports.

These strategies are then operationalized by the Dynamic Cyber Deception (DCD) module through structured deception playbooks, where each recommended technique is mapped to a corresponding deception tactic based on its category (Detection, Misdirection, Engagement, or Response) using a predefined association (as shown in Figure 2). This ensures that high-level strategic objectives are consistently translated into actionable technical deployments within the playbooks.

Continuous telemetry feedback informs both the Mission Impact Assessment (MIA) and future refinement of deception schemes, ensuring that the defense posture remains contextually relevant, adaptive, and aligned with mission objectives. The modular architecture enables flexibility, human-in-the-loop operation during early stages, and supports incremental evolution toward full AI-orchestrated cyber deception.

3.1 Information Collection

Building on the principles of dynamic risk assessment, the *Information Collection* module systematically aggregates environment-specific data, including Common Platform Enumeration (CPE) identifiers, Common Vulnerabilities and Exposures (CVE) vulnerabilities, and Exploit Prediction Scoring System (EPSS) exploit likelihood scores, to inform targeted cyber deception strategies. The model's operation begins with the identification of all relevant assets within the target environment. For each identified asset, the model determines the corresponding CPEs, which are then matched against the National Vulnerability Database (NVD) [2] to extract associated vulnerability data in the form of CVE identifiers. These CVE-IDs are subsequently used to retrieve EPSS [1] scores. The collected EPSS scores serve as input to a Bayesian network, which enables the model to dynamically and proactively estimate the likelihood of exploitation, producing quantitative threat assessments tailored to the specific environment [6].

In the subsequent phase, the model employs a sequential approach to identify related Common Weakness Enumerations (CWEs). These CWEs are then used to detect associated Common Attack Pattern Enumerations and Classifications (CAPECs). It is important to note that a single CVE may map to multiple CWEs; in such cases, the model extracts and analyzes each identified CWE. Similarly, each CWE may be associated with multiple CAPECs, all of which are collected by our model for further analysis. These identified CAPECs, along with their associated descriptions, are then provided as input to the Decision and Policy Engine (DPE) through the DRA module.

3.2 Dynamic Risk Assessment module

The *Dynamic Risk Assessment* module is responsible for integrating risk-related information (e.g., threat scores and impact levels) with target-specific contextual data, such as network topology. This integration is performed using a Bayesian

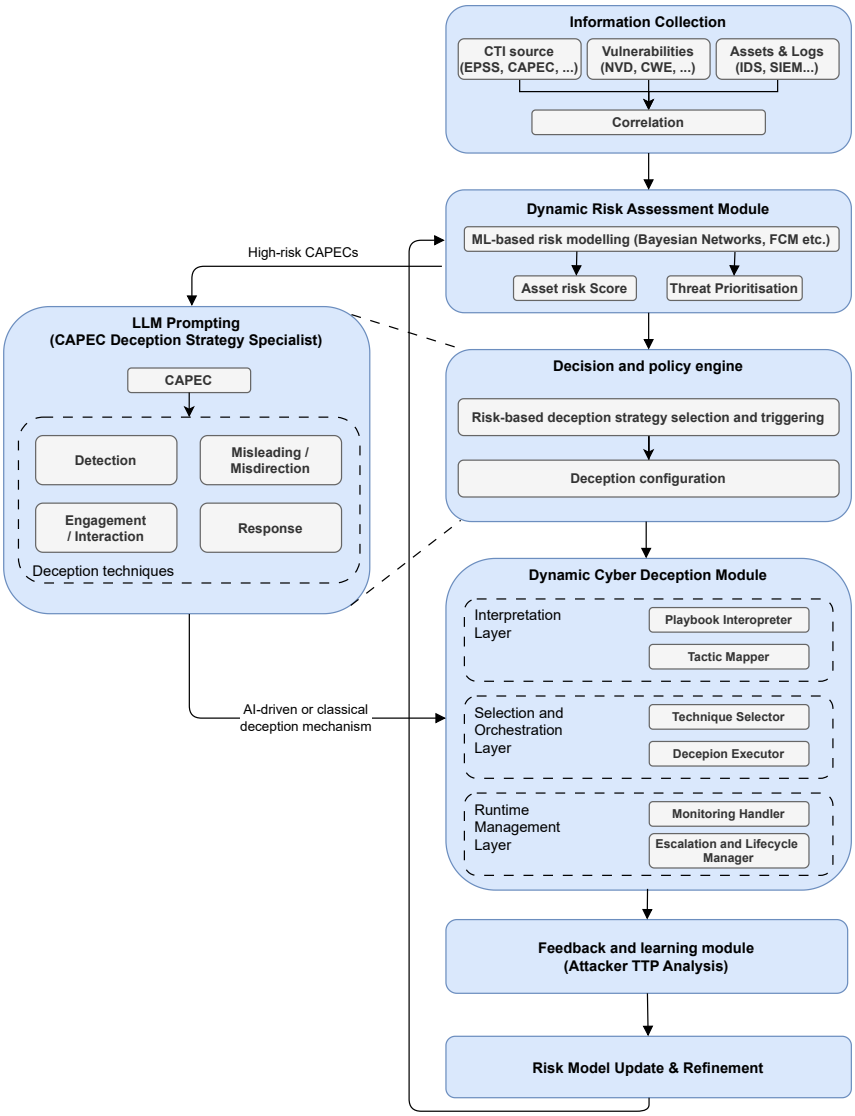

Fig. 1. Proposed framework for risk-driven cyber deception

Network (BN) to generate quantitative risk estimates tailored to the operational environment. It is important to note that this approach is one of several possible methods for Dynamic Risk Assessment, chosen here based on the authors' previous work and expertise in Bayesian Network modelling.

Bayesian Networks (BNs) have been extensively applied in cybersecurity for dynamic risk assessment due to their ability to capture probabilistic relationships and update threat estimations as new data becomes available [6,7]. A BN is defined as $N = \{G, P\}$, where $G = \{V, E\}$ is a Directed Acyclic Graph (DAG) comprising nodes V and edges E, and P represents conditional probability distributions [13,14,19].

In our framework, we define two node types: (T) *threat nodes*, representing CVEs that may exploit system vulnerabilities, and (A) *asset nodes*, representing systems potentially impacted. Rather than relying on subjective expert input [18], we derive conditional probabilities from EPSS scores and CAPEC data, which are widely accepted in the cybersecurity domain.

Network structure reflects possible attack paths informed by asset connectivity and threat propagation. To define node logic, we adopt the AND/OR gate approach from [20]. OR gates are used when any threat can compromise an asset or an asset can be reached by many connected assets, while AND gates require multiple conditions, such as threat presence and previous asset compromise. The conditional probabilities are computed as:

$$P_d = 1 - \prod_{i=1}^{n} \left(1 - P_{(i)}\right) \tag{1}$$

$$P_c = \prod_{i=1}^{n} P(i) \tag{2}$$

This setup enables dynamic threat updates and risk calculation using the standard formulation [15]:

$$Risk = \sum_{i} P(A_i) \times S(A_i) \tag{3}$$

Here, $P(A_i)$ denotes the posterior threat probability for asset A_i, and $S(A_i)$ is its impact score. Impact levels reflect service degradation (Table 1), though the model can accommodate alternative schemes (e.g., confidentiality-integrity and availability (CIA) or financial metrics).

The primary output of the DRA Module is a set of quantitative risk scores assigned to each asset in the network. These scores capture both the estimated likelihood of exploitation and the potential impact of a successful attack, supporting informed and prioritized decision-making for cyber defence planning. In addition, this module also supplies the DPE with the relevant CAPECs identified by the information collection component.

Table 1. Impact Scale

Description	Impact Score
All services operational	0
Most services operational	1
Some services operational	2
No services operational	3

3.3 Decision and Policy Engine

The *Decision and Policy Engine* serves as the semantic interpretation layer of the framework, responsible for transforming structured threat intelligence, particularly MITRE CAPEC entries identified by the DRA module, into actionable deception strategies. At the heart of the DPE is an LLM specifically configured and prompted to act as a "CAPEC Deception Strategy Specialist" [3]. This specialized agent is designed to analyze a given CAPEC, understand its attack mechanics, and generate detailed, creative, and technically precise deception-based countermeasures.

Workflow and Output Generation Upon receiving a high-risk CAPEC identifier and its associated context from the DRA module, the DPE invokes its specialized LLM agent. The agent processes this input based on its predefined system prompt, which guides it to first identify the specific CAPEC attack pattern, then provide an ultra-concise technical snapshot of the attack and its key enabling factors to ground the deception strategy, and ultimately generate a comprehensive Deception Strategy Report. This report is the primary output of the DPE and directly addresses the need for specific, actionable guidance.

The Deception Strategy Report is structured to include several key components, directly aligning with the capabilities of the "CAPEC Deception Strategy Specialist" agent. It details a set of recommended deception techniques which are specific deception tools and methods, such as credential honeytokens, fake vulnerable service honeypots, and deceptive API endpoints, directly relevant to countering the mechanics of the input CAPEC. This "set" of techniques is generated by the LLM agent based on its specialized knowledge of deception and the CAPEC landscape.

Furthermore, for every recommended deception technique, the report provides clearly defined strategic objectives, such as early detection of reconnaissance, misdirection of the attacker towards benign decoys, containment of malicious activity, or collection of attacker TTPs. These objectives are defined by the LLM agent itself, tailored to the specific CAPEC and the chosen deception tactic.

The agent also provides implementation guidance, offering practical advice on the optimal placement of these deception elements within the target environment (e.g., specific network segments, application layers, configuration files) and

detailing crucial monitoring requirements (e.g., "alert on any access attempt to the honeytoken", "log all traffic interacting with the honeypot IP").

This guidance forms the strategic basis for the more detailed operational playbooks. Crucially, the report also establishes the basis for trigger logic; while the LLM agent does not generate executable Security Information and Event Management (SIEM) rules or complex alert conditions directly, its specification of "key monitoring points" and expected attacker interactions provides the necessary information for the subsequent DCD module to establish explicit trigger logic for high-confidence alerts or automated escalations. For instance, if the agent suggests monitoring a specific fake credential, the DCD module can then implement a trigger to alert when that credential is used.

Example Output of the "CAPEC Deception Strategy Specialist" Agent To illustrate the output, consider an input of **CAPEC-126: Path Traversal**. The DPE's LLM agent might produce a Deception Strategy Report similar to the following (abbreviated):

```
### CAPEC Deception Strategy Report

**Attack Pattern:** CAPEC-126: Path Traversal (aka Directory
    ↪ Traversal)

**Threat Snapshot:** An attacker manipulates input parameters
    ↪  (e.g., in a URL or form field) with sequences like
    ↪ '../' to navigate outside the intended directory and
    ↪ access restricted files or directories on the server.

**Key Enabling Factor(s):**
*   Insufficient input validation and sanitization of user-
    ↪ supplied file paths.
*   Web server misconfiguration allowing access beyond the
    ↪ web root.

---

**Deception Strategy:**

**1. Technique:** Decoy Directory/File System
    *   **Details:** Create a set of directories and files
        ↪ within the web server's accessible paths that
        ↪ mimic sensitive system locations (e.g., '/etc/
        ↪ passwd-decoy', '/var/log/secret_system.log.decoy',
        ↪  '../backup_config.ini.decoy'). These are not real
        ↪  sensitive files but are designed to attract path
        ↪ traversal attempts.
    *   **Strategic Goal:** Detect and identify attackers
        ↪ attempting path traversal by luring them to access
        ↪  these specific decoy files/directories.
```

```
        ↪ Successful access to a decoy is a high-fidelity
        ↪ indicator of malicious intent.
    *   **Implementation & Monitoring:**
        *   Placement: Strategically place links or
            ↪ references to these decoys in less obvious
            ↪ parts of the web application, or make them
            ↪ discoverable if an attacker successfully
            ↪ traverses one level up.
        *   Monitoring: Monitor all access attempts to these
            ↪ specific decoy files and directories. Log the
            ↪ source IP, timestamp, and the exact path
            ↪ requested.

**2. Technique:** Path Traversal Honeypot Listener
    *   **Details:** Configure a specific web application
        ↪ endpoint or a dedicated lightweight honeypot
        ↪ service that intentionally appears vulnerable to
        ↪ path traversal. When an attacker attempts to use
        ↪ `../` sequences against this endpoint, instead of
        ↪ returning a real file or an error, it logs the
        ↪ attempt and can optionally return a fake directory
        ↪  listing or a "file" containing a honeytoken.
    *   **Strategic Goal:** Actively solicit and capture path
        ↪  traversal attempts, gather intelligence on
        ↪ attacker TTPs (what files they are looking for),
        ↪ and provide early warning of such reconnaissance.
    *   **Implementation & Monitoring:**
        *   Placement: Can be a hidden endpoint on an
            ↪ existing web server or a separate low-
            ↪ interaction honeypot.
        *   Monitoring: Log all incoming requests to this
            ↪ listener, especially those containing path
            ↪ traversal sequences. If a honeytoken is
            ↪ returned, monitor for its activation.

---

**Brief Contextual Notes:**
*   **Attacker Objective Hint:** Often used for information
    ↪ gathering (e.g., finding configuration files,
    ↪ credentials) or as a precursor to further exploitation
    ↪ .
*   **Deception Complexity:** (Technique 1: Simple to
    ↪ Moderate), (Technique 2: Moderate)
```

This structured "Deception Strategy Report" serves as input to the DCD Module. It is the blueprint for constructing and operationalizing executable deception playbooks. The guidance provided by the DPE ensures that the deception mechanisms deployed are semantically rich, aligned with specific threats

(CAPECs), and contribute directly to the strategic objectives of detection, misdirection, engagement, or response.

The DPE, powered by a specialized LLM agent acting as a "CAPEC Deception Strategy Specialist", transforms structured threat intelligence into dynamic and semantically rich deception strategies. By interpreting CAPEC data, the engine generates detailed Deception Strategy Reports (Deception Scheme) that recommend specific deception techniques, associate them with strategic objectives, and provide implementation and monitoring guidance. This output serves as the direct input for the DCD module, enabling the creation of context-aware, risk-aligned deception playbooks tailored to evolving threats and operational constraints.

3.4 Dynamic Cyber Deception module

The purpose of the *Dynamic Cyber Deception* module is to execute deception playbooks based on recommendations provided by the DPE. Acting as a bridge between AI-driven decision modules and real-time deceptive actions, playbooks define when and how deception should be activated in response to specific attacker behaviors or risk levels, as well as the intended outcomes, i.e. whether to detect, mislead, engage, or respond to cyber attackers according to the deception categories identified in the framework (Fig 1). In authors' views, the DCD module follows a human-based implementation where operators are responsible for interpreting high-level deception schemes/deception strategy reports, mapping tactics to executable techniques, deploying deception components, configuring monitoring, and managing the full lifecycle of each playbook. This manual workflow emphasizes the need for a structured transition plan toward automated execution in the future. A deception playbook is not merely an execution recipe, but a structured operational plan that translates the high-level deception scheme (from the DPE) into: tactical intent (why it is being used), technical deployment (how it is executed) and expected effects (what it is supposed to achieve).

The intented outcome is not inferred post-deployment but is embedded in the playbook structure to ensure consistent, goal-driven deception planning and evaluation. It instantiates both a "classical approach" - comprising, for instance, honey-X and MTD techniques - and an "AI-based approach" of adaptive and cognitive deception. Both technologies can coexist in a hybrid deception architecture. Human operators are responsible for interpreting and executing this specification, which highlights the need for a structured transition to automated orchestration as system maturity evolves. The DCD module may run immediate deception via a rapid response path and generates outputs shared with a feedback loop for continuous improvement. It is the "executor and tuner" of deception based on the input received from previous modules. Here is how the module can be functionally decomposed into three logical layers containing each one two submodules and services (as shown in Fig 2) - in a structured way:

- **The Interpretation Layer** serves as the central coordination of two submodules: playbook interpreter and tactic mapper. The playbook interpreter

parses DPE inputs, validates deception schema, and instantiates a runtime deception playbook while the tactic mapper maps the deception category (e.g., misleading) to a corresponding tactic (e.g., decoying or camouflage) using a predefined lookup table. The category-to-tactic mapping is a semantic grounding mechanism that refines high-level deception objectives into specific tactical intents. The original set of deception tactics proposed in the CYDEC framework [4] was designed to capture high-level cognitive strategies applicable to both offensive and defensive deception contexts. While valuable as a unifying taxonomy of intent, these tactics were not explicitly structured for implementation within adaptive, mission-aware cyber defence systems. In this work, the authors reinterpret and expand these tactics to align with four operationally distinct categories of deception (Fig 2), targeting *detection*, *misdirection*, *engagement*, or *response* and thus, creating a practical bridge between strategic intent and automated execution. This expansion not only enhances the applicability of existing taxonomies but also provides the foundation for integrating deception planning into AI-driven decision systems and dynamic orchestration modules.

- **The Selection and Orchestration Layer** comprises two submodules: a Technique selector and a Deception executor. The first submodule, analyses the list of techniques provided by the DPE and selects the one that best implements the mapped tactic (based on priority, resource availability, or historical effectiveness). The latter submodule deploys the selected deception technique using orchestration tools (e.g., honeypot deployment, DNS manipulation, AI-generated content).
- **The Runtime Management Layer** manages telemetry setup, tracks attacker behaviour, and governs the playbook lifecycle including escalation or shutdown. These functions are operationalised by a Monitoring Handler and an Escalation and Lifecycle Manager.

In essence, playbooks incorporate trigger logic, placement and duration parameters, telemetry hooks, and escalation rules as defined by the LLM-generated scheme. This alignment ensures that each deception instance is semantically consistent, mission-relevant, and operationally effective within the adaptive deception loop. Unlike SODA [21], which synthesizes deception playbooks through offline malware analysis, the proposed model constructs deception playbooks based on semantic interpretations of elements provided by an LLM-driven DPE. SODA's playbooks are primarily malware-specific and centered around ploy-level API manipulation aligned with static deception strategies (e.g., FakeSuccess, FakeFailure), whereas our approach defines playbooks as modular execution plans guided by high-level strategic objectives (e.g., detection, misdirection, engagement, response), linked to mapped tactics and techniques.

Together, the above-mentioned submodules form the active operational layer of the dynamic deception architecture, seamlessly integrating deception deployment with situational understanding. In the authors' opinion, implementation of the DCD module will include human oversight in key subcomponents. Specifically, the Technique Selector and Deception Executor submodules are envisaged

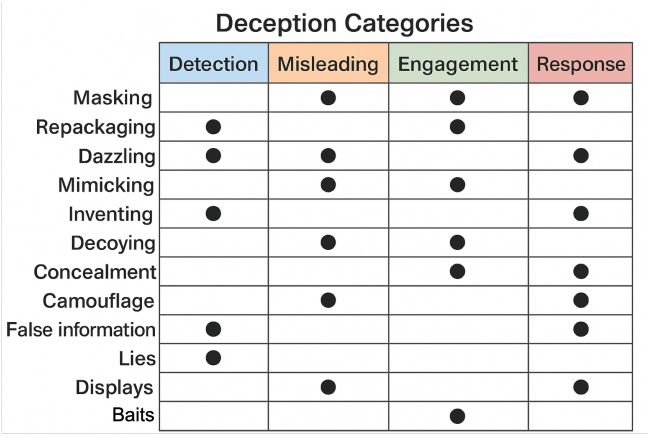

Fig. 2. Mapping of CYDEC tactics [4] with deception categories

to be operated manually or semi-automatically by human analysts during the proof-of-concept phase. This design choice ensures flexibility and interpretability in the early stages of system deployment, allowing expert judgment to guide the selection of deception techniques and validate operational execution in complex environments [17]. Furthermore, this human-in-the-loop approach enables iterative refinement of category-tactic-technique mappings and supports the safe evaluation of deception strategies prior to some type of possible automation. In production settings, these submodules are expected to evolve into policy-driven, AI-assisted orchestration services, but their manual control at this stage reflects a balance between experimental rigor, operational safety, and the incremental maturity of deception technology.

3.5 Feedback and improvement of the model

The *Feedback Processor* serves as a critical analytical component within the deception architecture, transforming raw telemetry and attacker interaction data into actionable insights for system adaptation and learning. It continuously ingests data streams from the Monitoring Handler, including command sequences, engagement durations, and deception avoidance attempts. Utilizing advanced analytics, the processor assesses the effectiveness of deception activities, measuring indicators such as attacker persistence, confusion, and eventual disengagement. This analysis produces deception performance metrics that are structured and prioritized according to their relevance to ongoing operations and overall mission objectives. These metrics are then fed both to the MIA module, where they contribute to dynamic impact evaluation, and to the DPE module, facilitating continuous refinement of deception playbooks and orchestration policies based on empirical outcomes. Additionally, it has the potential of executing an adaptive tuning based on dynamic risk and MIA [9] - the latter being an external

Table 2. Functional overview of Deception module subcomponents

Submodule	Purpose
Playbook Interpreter	Parses DPE inputs, validates schema, and instantiates a runtime deception playbook
Tactic Mapper	Maps the deception category to a corresponding tactic using a lookup table
Technique Selector	From the list of techniques provided by the DPE, selects the one that best implements the mapped tactic (based on priority, resource availability, or historical effectiveness)
Deception Executor	Deploys the selected deception technique using orchestration tools (e.g., honeypot deployment, DNS manipulation, AI-generated content)
Monitoring Handler	Installs required telemetry hooks as defined in the playbook and routes data to the Feedback Processor and MIA modules
Escalation and Lifecycle Manager	Monitors playbook triggers and conditions for escalation, timeout, or deactivation, and coordinates any runtime transitions

source of information. In a cyber situational awareness capability, MIA evaluates how ongoing cyber events affect mission-critical capabilities. In doing so, the Feedback Processor closes the deception loop, ensuring that future engagements benefit from accumulated knowledge and tactical optimization.

4 Discussion

In the implementation of the framework provided in Fig. 1, some shortfalls may appear which are worth analysing. As mentioned in section 3, LLM outputs map directly to deception categories (detect, mislead, engage, respond). Deception tactics associated with above-mentioned categories could be predictable or rigidly tied to CAPEC mappings to cope with a rapid adaption of cyber attackers's techniques. Therefore, Deception playbooks need to evolve dynamically, not just trigger based on category. Another possible limitation is that the DCD module produces attacker reactions which are captured to feed a MIA. MIA consumes information from deception activities and attacker reactions to provide its measurements. The proposed future work for the proposed architecture shall show how MIA influences upstream deception or LLM recommendations. DPE module is not directly fed with real-time attacker response data and it would affect the ability for the LLM-driven mechanism to refine recommendations based on deception effectiveness. Additionally, the modular approach and involvement of LLMs and agents may introduce processing latency. High-speed attacks (e.g. ransomware, data exfiltration) require faster LLM-based decision and deception deployment or at least, the possibility to find a compromise between effectiveness and speed. Moreover, explicit adversarial resistance is envisaged as a future work. Attackers may attempt to detect and evade LLM-driven deception systems. LLM and deception mechanisms may be vulnerable to: adversarial prompting, decep-

tion environment fingerprinting, or abuse of response deception to gain information. Successfully fingerprinting deception environments allows adversaries to bypass traps, reducing the effectiveness of cyber deception strategies.

The implementation of the DCD module requires careful orchestration of modular subcomponents capable of deploying and adapting deception strategies with human intervention in alignment with evolving threat conditions and mission priorities. Central to this is the integration of deception playbooks, which standardize the execution of deception schemes, ranging from detection and misdirection to environment manipulation and response deception, based on structured outputs from the DPE. Key aspects include the use of containerized microservices for flexible deployment, real-time telemetry collection for attacker interaction analysis, and a feedback mechanism to inform future deception strategies. At the framework level, seamless interoperability between the the DRA process, and the DPE and DCD modules is critical to ensure context-aware and risk-aligned defensive actions. To ensure a coherent and logically structured operationalization of the framework, overlaps or inconsistencies between the inputs and outputs of the DPE and the DCD modules must be minimized. Clear interface definitions and modular boundaries are essential to maintain consistency, traceability, and semantic alignment across the system components. The definition of deception intensity within the DCD module must consider both the availability of operational resources and the assessed threat level. This intensity should be proportionally aligned with the threat actor's motivation and skills, enabling a calibrated response that avoids unnecessary system overhead. Furthermore, maintaining low-latency communication, ensuring scalability across distributed environments, and safeguarding against deception environment fingerprinting are essential to achieving a resilient, adaptive, and mission-aware cyber defence capability.

5 Conclusions

This work introduces a novel cyber defence framework that integrates DRA with AI-assisted cyber deception, aiming to enhance mission resilience through adaptive, risk-aligned defensive actions. Leveraging Large Language Models to interpret CAPEC-aligned threat intelligence, the system generates semantically rich deception strategies that guide deployment of structured deception playbooks.

The proposed framework establishes a modular, closed-loop architecture linking risk analysis, strategic reasoning, deception orchestration, and feedback-driven adaptation. While the current work focuses on the architectural foundation, future research will focus on the detailed development and operational refinement of the framework components, including the integration of feedback mechanisms, advanced deception environment and playbooks management, and optimisation of AI-driven strategy generation. Before DCD deployment of playbooks, alternative tactics and techniques analysis can be effectively supported through controlled testing environments or cyber ranges, allowing empirical evaluation of deception strategies and identification of the most effective solution for

a given threat scenario. By bridging dynamic risk evaluation with proactive, context-aware cyber deception, this framework lays the groundwork for the next generation of intelligent and mission-driven cyber defence systems where cyber deception is an integral part of it.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Exploit Prediction Scoring System (EPSS). www.first.org/epss/, accessed on August 13, 2024
2. National Vulnerability Database (NVD). www.nvd.nist.gov/, accessed on August 10, 2024
3. Batzolis, E.: Github - lefteris-b/system_prompts_for_deception_agents: System prompts for ai cyber deception agents. https://github.com/Lefteris-B/System_prompts_for_Deception_Agents (May 2025), accessed: 2025-05-12
4. Beltrán López, P., Gil Pérez, M., Nespoli, P.: Cyber Deception: State of the art, Trends and Open challenges (Sep 2024). https://doi.org/10.48550/arXiv.2409.07194, http://arxiv.org/abs/2409.07194, arXiv:2409.07194 [cs]
5. Cheimonidis, P., Rantos, K.: Dynamic risk assessment in cybersecurity: A systematic literature review. Future Internet **15**(10), 324 (2023)
6. Cheimonidis, P., Rantos, K.: A dynamic risk assessment and mitigation model. Applied Sciences **15**(4), 2171 (2025)
7. Cheimonidis, P., Rantos, K.: A novel proactive and dynamic cyber risk assessment methodology. Computers & Security **154**, 104439 (2025)
8. De Faveri, C., Moreira, A., Amaral, V.: Multi-paradigm deception modeling for cyber defense. Journal of Systems and Software **141**, 32 – 51 (2018). https://doi.org/10.1016/j.jss.2018.03.031
9. Grimaila, M.R., Mills, R.F., Fortson, L.W.: Improving the cyber incident mission impact assessment (cimia) process. In: Proceedings of the 4th Annual Workshop on Cyber Security and Information Intelligence Research: Developing Strategies to Meet the Cyber Security and Information Intelligence Challenges Ahead. CSIIRW '08, Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1413140.1413177
10. Huang, L., Zhu, Q.: A dynamic games approach to proactive defense strategies against advanced persistent threats in cyber-physical systems. Computers and Security **89** (2020). https://doi.org/10.1016/j.cose.2019.101660
11. Islam, M.M., Al-Shaer, E.: Active deception framework: An extensible development environment for adaptive cyber deception. In: 2020 IEEE Secure Development (SecDev). pp. 41–48. IEEE (2020)
12. Javadpour, A., Ja'fari, F., Taleb, T., Shojafar, M., Benzaïd, C.: A comprehensive survey on cyber deception techniques to improve honeypot performance. Computers & Security p. 103792 (2024)
13. Jensen, F., Nielsen, T.: Bayesian Networks and Decision Graphs. Springer, 2nd edn. (2007)
14. Johnson, P., Lagerström, R., Ekstedt, M., Franke, U.: Can the common vulnerability scoring system be trusted? a bayesian analysis. IEEE Transactions on Dependable and Secure Computing **15**(6), 1002–1015 (Nov 2018)

15. Kaplan, S., Garrick, B.J.: On the quantitative definition of risk. Risk analysis **1**(1), 11–27 (1981)
16. Li, H., Guo, Y., Sun, P., Wang, Y., Huo, S.: An optimal defensive deception framework for the container-based cloud with deep reinforcement learning. IET Information Security **16**(3), 178 – 192 (2022). https://doi.org/10.1049/ise2.12050
17. Llopis Sanchez, S., Lopes Antunes, D.: Operation assessment in cyberspace: Understanding the effects of cyber deception. In: Proceedings of the 19th International Conference on Availability, Reliability and Security. ARES '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3664476.3672355
18. Mkrtchyan, L., Podofillini, L., Dang, V.: Methods for building conditional probability tables of bayesian belief networks from limited judgment: An evaluation for human reliability application. Reliab. Eng. Syst. Saf. **151**, 93–112 (2016). https://doi.org/10.1016/j.ress.2016.01.004
19. Peng, Y., Huang, K., Tu, W., Zhou, C.: A model-data integrated cyber security risk assessment method for industrial control systems. In: 2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS). pp. 344–349. IEEE (2018)
20. Poolsappasit, N., Dewri, R., Ray, I.: Dynamic security risk management using bayesian attack graphs. IEEE Transactions on Dependable and Secure Computing **9**(1), 61–74 (Jan 2012)
21. Sajid, M.S.I., Wei, J., Abdeen, B., Al-Shaer, E., Islam, M.M., Diong, W., Khan, L.: Soda: A system for cyber deception orchestration and automation. In: Proceedings of the 37th Annual Computer Security Applications Conference. p. 675–689. ACSAC '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3485832.3485918
22. Sengupta, S., Chowdhary, A., Sabur, A., Alshamrani, A., Huang, D., Kambhampati, S.: A survey of moving target defenses for network security. IEEE Communications Surveys and Tutorials **22**(3), 1909 – 1941 (2020). https://doi.org/10.1109/COMST.2020.2982955
23. Wang, R., Yang, C., Deng, X., Zhou, Y., Liu, Y., Tian, Z.: Turn the tables: Proactive deception defense decision-making based on bayesian attack graphs and stackelberg games. Neurocomputing **638** (2025). https://doi.org/10.1016/j.neucom.2025.130139