Issues in Information Systems
Volume 27, Issue 2, pp. 122-136, 2026

DOI: https://doi.org/10.48009/2_iis_2026_109

AI purple teaming as experiential learning: A framework for cross-functional security expertise

Kathleen Gibson, *Marymount University, kathleen_gibson@marymount.edu*

Abstract

The rapid deployment of artificial intelligence (AI) systems across enterprises has created an urgent challenge: developing security expertise across the cross-functional teams that build, deploy, and use them. AI purple teaming is emerging as a collaborative approach that combines adversarial testing with detection validation to surface and address AI security and safety risks. However, its dynamics as an organizational learning process remain under-theorized. Building on experiential learning theory and reflective practice, this paper conceptualizes AI purple teaming as a collaborative organizational learning practice in which the experiential learning cycle unfolds through interactions among participants with different organizational roles. Throughout the scoping, testing, and reporting phases of an AI purple team engagement, security professionals, developers, and business partners build a shared understanding of AI risk. The framework proposes that this understanding complements the engagement's testing and validation outputs, yielding benefits beyond those of conventional security approaches. The paper contributes this framework, eight design principles for engagements that cultivate cross-functional security expertise, and a positioning of AI purple teaming within AI risk management.

Keywords: AI purple teaming, experiential learning, reflective practice, AI security, AI red teaming, organizational learning

Introduction

In a typical AI purple team engagement, security professionals work through scoping conversations with developers and business partners to identify what to test, which threats matter, and which assumptions to probe. After testing concludes, they reconvene to share findings, prioritize risks, and provide remediation recommendations. Between these conversations, the red team (offensive security) runs adversarial tests while the blue team (defensive security) monitors the alerting pipeline and logs. Across these phases, participants from different functions observe the results of attacks in real time and how security controls respond or fail to respond in their systems. The framework proposes that by collaborating on engagements and observing how others reason about their methodologies and actions, participants can develop a shared understanding of AI risk that is less likely to emerge from policy documents or training alone. Engagements can generate knowledge that had not existed in codified form (Nonaka, 1994).

This paper examines the learning process, drawing on experiential learning theory and reflective practice to develop a framework for AI purple teaming. The rapid pace of AI deployment in enterprise contexts makes this urgent. By one estimate (Stanford HAI, 2026), generative AI reached 53% population adoption within three years, faster than personal computers or the internet. Security expertise must develop quickly across the security, development, and business functions involved in AI system development. Experiential learning theory has been widely applied across professional education and practice contexts, including

medicine, higher education, business, and engineering (Kolb & Kolb, 2017; Schön, 1983), yet its application to AI security practices remains under-theorized.

Recent research has examined key areas of AI security testing. Gillespie et al. (2026) frame AI red teaming as a sociotechnical problem that requires attention to values, labor, and harms. Anthropic researchers describe a whole-system approach to red teaming language models to identify and mitigate vulnerabilities and potential harms in AI systems (Ganguli et al., 2022). Majumdar et al. (2025) propose a system-level framework for red teaming across the AI development lifecycle. Microsoft researchers conducted extensive red teaming of Generative AI products, revealing the sociotechnical complexity of AI security evaluation, in which human judgment, collaboration, and adaptive learning remain central to identifying and addressing emerging risks (Bullwinkel et al., 2025). These contributions establish AI red teaming as a substantive area of research. AI purple teaming, however, has received less attention as a distinct practice with its own organizational dynamics. The learning effects of these engagements on the cross-functional participants who carry them out remain underexamined.

The framework conceptualizes AI purple teaming as a collaborative, experiential learning practice in which security professionals, developers, and business partners build cross-functional security expertise through the conversations, artifacts, and reasoning generated by testing engagements. Throughout the scoping, testing, and reporting phases, the experiential learning cycle operates among participants rather than within a single learner, as each participant models their reasoning for others and learns from the knowledge and context others bring.

The paper proceeds as follows. The literature review builds the theoretical foundation in experiential learning theory and reflective practice and highlights AI purple teaming as an emerging practice. The methodology section outlines the conceptual approach and scope, while the framework section details the experiential learning dynamics across the engagement phases. The discussion explores implications for practice and research, including design principles for engagements. The limitations section clarifies the boundaries of analysis, and the conclusion synthesizes the central contribution and its significance.

Background and Literature Review

Experiential Learning: Foundations and Applications

Experiential learning theory (ELT) was developed to provide a theoretical foundation for learning through experience, drawing on the foundational work of John Dewey (1938) in experiential education and Kurt Lewin (1951) in action research and group dynamics (Kolb, 2015; Kolb & Kolb, 2017). Kolb and Kolb (2005, 2017) integrated these traditions into a holistic learning model that conceptualizes learning as a continuous cycle of "thinking, feeling, perceiving, and behaving" (Kolb & Kolb, 2005, p. 194). Learning is the process by which knowledge emerges from the transformation of experience, shaped by how individuals grasp and transition that experience (Kolb & Kolb, 2005; Kolb, 2015). The ELT cycle begins with concrete experience, an active engagement with a task, problem, or situation. This is followed by reflective observation, in which the learner reflects on what has occurred. The learner then moves to abstract conceptualization, developing ideas or general principles from those reflections, transforming experience into knowledge. The cycle concludes with active experimentation, in which these insights are applied in practice, generating new experiences and restarting the process. While the cycle describes how individuals transform experience into knowledge, Kolb and Kolb (2017) emphasize that it also operates across groups, organizations, and society. On this view, collaborative settings can produce learning that individual participants are less likely to generate on their own.

Schön's (1983) work on reflective practice extends this foundation to professional contexts, emphasizing how practitioners learn through reflection-in-action, by questioning and adjusting their approach as they engage with real problems, and through reflection-on-action, by looking back on past practice to draw lessons that inform subsequent work. Schön's framework also addresses organizational learning, theorizing that groups of practitioners learn collectively through inquiry into their shared theories of action and through dialogue that surfaces and tests underlying assumptions. In organizational settings, this reflection unfolds through discussion and collaboration with others who bring different expertise to shared problems, with practitioners learning not only from their own experience but also by observing how colleagues frame, reason about, and act on the same situation.

Building on this theoretical foundation, experiential learning has been applied in cybersecurity education to bridge the gap between conceptual knowledge and practical skills. Nairan et al. (2025) describe a course model grounded in ELT that integrates experiential activities, such as labs, gamified simulations, and immersive training environments, to engage students throughout the learning cycle. Their approach emphasizes reflection and conceptual understanding to complement technical skill development, demonstrating how ELT can be operationalized within domain-specific curricula to enhance student learning outcomes.

In the cybersecurity industry, experiential methods are typically collaborative. Practitioners work together on systems and tooling, design and conduct red-team scenarios and incident-response drills, and participate in gamified learning settings such as capture-the-flag contests and tabletop exercises. These practices share a common foundation in the ELT cycle. Practitioners engage in concrete experiences, reflect on what happened, abstract general principles, and apply those insights in subsequent practice. However, the cycle unfolds collectively across organizational roles rather than through an individual learner's experiences. Participants observe how others approach the same problem, surface assumptions through dialogue, and refine their own reasoning in response. While these practices are well established in cybersecurity, they have not been systematically examined through the lens of experiential learning theory, and their application to AI purple teaming as a specific organizational practice remains under-researched.

What distinguishes AI purple teaming from these established practices is the nature of the learning object. Unlike traditional exercises that test stable, deterministic systems with vulnerabilities that persist until fixed, this object is dynamic and constantly changing. Its behavior is based on learned patterns; it evolves as models are updated and adversarial techniques advance, and it acts more agentically by leveraging tools and performing multi-step actions. These properties reshape the experiential learning cycle, altering how lessons abstracted in one engagement carry over to the next.

AI Red Teaming

As AI deployment rapidly expands, AI security and safety incidents have risen sharply, with documented incidents increasing from 233 in 2024 to 362 in 2025 (Stanford HAI, 2026), creating urgent demand for evaluation practices that can identify and mitigate risks before deployment. AI red teaming has emerged as a key practice for evaluating the security and safety of AI systems through adversarial testing (Bullwinkel et al., 2025; Gillespie et al., 2026). It applies the practice of red teaming, systematic adversarial testing in which testers emulate real-world attacks to identify vulnerabilities, to AI systems by generating inputs designed to expose harmful or unsafe behaviors (Perez et al., 2022). Ganguli et al. (2022) report that models incorporating red team techniques can reduce harmful outputs. Frameworks such as MITRE Corporation's (n.d.-a) Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS), the OWASP Foundation’s (2024) Top 10 for Large Language Models, and the OWASP Foundation’s (2025) Top 10 for Agentic AI Applications provide structured representations of risks and attack techniques against AI systems, enabling exploration of vulnerabilities. The National Institute of Standards and Technology

(NIST, 2023) AI Risk Management Framework (RMF) positions adversarial testing as one component of broader organizational risk management practice, organized around the functions of Govern, Map, Measure, and Manage. The AI RMF establishes what organizations should do to manage AI risk, but provides limited theoretical attention to how the cross-functional security expertise required to enact these functions develops in practice. These approaches focus on what to test, which risks to manage, and which frameworks to apply, with comparatively less attention to the structured learning processes that emerge among cross-functional participants in adversarial engagements.

Purple Teaming

Purple teaming is a collaborative approach in information security that integrates red (offensive testing) and blue (defensive monitoring, detection, and incident response) functions to strengthen an organization's security posture through coordinated exercises and assessments (SCYTHE, 2026). It is an iterative process in which offensive and defensive improvements are continually refined, creating a feedback loop that helps strengthen security over time. A typical purple team includes roles focused on collecting threat intelligence, conducting threat modeling, designing attack scenarios, simulating adversarial activity, monitoring detection and logs, and enhancing security controls to reduce risk. Attack scenarios are often mapped to frameworks such as the MITRE ATT&CK matrix and the Cyber Kill Chain (Lockheed Martin, n.d.; MITRE Corporation, n.d.-b; Routin et al., 2022; SCYTHE, 2026). While the red and blue functions form the core of a purple team, engagements typically extend beyond the security team, drawing in developers and stakeholders whose systems are being tested and whose participation shapes the exercise.

Purple teaming has evolved largely through practitioner-driven methodologies, while its organizational dynamics and adoption patterns remain comparatively underexamined in academic research. Nevertheless, practitioner literature from organizations such as SANS Institute (Vandeleur, 2024) and SCYTHE (2026) offers detailed frameworks for designing and applying engagement strategies across diverse security domains. Although scholarly research on purple teaming is beginning to formalize its core concepts (Van Buggenhout, 2024), limited attention has been given to purple teaming as an organizational learning process, particularly how engagements develop cross-functional expertise and facilitate learning through collaborative participation. Furthermore, AI purple teaming, as a growing extension of these practices, has received little academic attention as a distinct organizational phenomenon.

AI Purple Teaming

The extension of purple teaming into AI is a relatively recent but rapidly evolving area of practitioner research. Some implementations focus heavily on automation, such as Meta's (2023) Purple Llama project and Lasso Security's agentic purple teaming, which combines adversarial testing with real-time defenses through autonomous systems (Dror, 2025). Others center on human collaborative engagements, which this paper focuses on and that remain underexplored in existing research. These engagements typically begin with collaborative scoping discussions in which security professionals and developers define the testing scope, identify the most relevant threats, and establish assumptions about model behavior worth probing. Business partners often participate in these conversations as well, contributing operational context and helping shape acceptable risk assumptions.

During testing, the red team assesses the AI for vulnerabilities such as prompt injection, jailbreaks, and policy breaches (Sotiropoulos, 2024), while the blue team monitors logs and alerts to ensure that controls function correctly and periodically exchanges with developers as questions arise about system behavior. The process typically concludes with a reporting phase in which findings are presented, prioritized, and translated into remediation recommendations through collaborative dialogue among the participating functions. This study examines these engagements through the lens of experiential learning theory and

reflective practice, addressing the gap in the application of ELT to security practice and exploring how this collaboration develops cross-functional security expertise among all participants.

Main Contribution of the Paper

This paper advances a framework that positions AI purple teaming as collaborative organizational learning in enterprise settings. Drawing on practitioner experience in AI security and safety engagements, experiential learning theory, and reflective practice, the analysis examines how security professionals, developers, and business partners build cross-functional security expertise across the scoping, testing, and reporting phases of an engagement. The framework extends experiential learning theory by showing how the cycle is distributed across organizational roles and how it orients toward the continuous reconstruction of risk models. It reframes AI purple teaming as a practice for developing security expertise, rather than as vulnerability testing alone. The paper contributes eight design principles for organizations structuring engagements and situates the practice within AI risk management.

Research Methodology

This study adopts a conceptual theory-building approach that iteratively integrates practitioner observations with established theoretical perspectives. The framework emerged from a reflective analysis of recurring patterns observed across AI red and purple team engagements in which the author participated directly or served in coordination roles. AI purple teaming, as addressed in this paper, encompasses both AI security concerns (adversarial threats to systems) and AI safety concerns (harmful behaviors of AI systems), recognizing that these issues overlap. Although the observations are not presented as systematic empirical data, they ground the framework's claims about how cross-functional learning unfolds in practice.

The literature review incorporated practitioner-oriented sources, including industry publications, vendor documentation (e.g., SCYTHE, SANS Institute), and professional frameworks (e.g., guidance from the OWASP Foundation and MITRE Corporation) to characterize AI purple teaming practices. Academic sources included peer-reviewed research and seminal works in information systems, cybersecurity (Nairan et al., 2025), and experiential and organizational learning (Argyris & Schön, 1996; Kolb, 2015; Kolb & Kolb, 2017; Nonaka, 1994; Schön, 1983; Wegner, 1987). The author analyzed observations from engagements by comparing patterns across cases that varied in system type, organizational context, and engagement scope, focusing on recurring dynamics and variations to reveal the contextual boundaries of these patterns. Practitioner observations of recurring engagement patterns were first identified, after which experiential learning theory and reflective practice were applied as analytical lenses. The framework was refined as the theoretical concepts highlighted, challenged, or extended the practitioner's reflective observations. Comparing across cases in this way mitigates, but does not eliminate, the single-practitioner bias inherent in an experience-grounded account. The framework reflects a single professional vantage point, and incorporating multiple observers and organizations is a promising direction for future research.

The framework's scope is bounded in four ways. First, it addresses AI purple teaming within organizations that develop and operate AI products, including enterprise AI applications such as coding and research assistants, as well as other agentic systems. Second, it focuses on the collaborative phases of AI purple team engagements (scoping, testing, and reporting) and excludes internal preparatory activities, such as adversary emulation planning and tooling selection. Third, it examines cross-functional learning among participants, including security professionals, developers, and business partners. It does not address learning confined entirely to internal security team activities, such as red team coordination or blue team work that does not

involve broader cross-functional participation. Fourth, the framework advances conceptual rather than empirical claims, grounded in practitioner experience and the existing literature, and is intended to complement established AI risk management frameworks.

The AI Purple Teaming Practice

This framework analyzes AI purple teaming as experiential learning shared across organizational roles. It applies the four modes of Kolb and Kolb's (2017) experiential learning theory (ELT) cycle to examine how cross-functional learning develops across engagement phases, with reflective practice enhancing the account of how these modes operate in professional settings. The units of analysis are as follows. Concrete experience (CE) is direct engagement with adversarial testing, threat scenarios, and system behavior. Reflective observation (RO) is participants' reflection on the experience and consideration of what occurred, both individually and in dialogue with others. Abstract conceptualization (AC) is participants' learning from the experience and the sharing of principles and mental models about AI risk. Active experimentation (AE) is participants' application of what they learned, both within the engagement and in subsequent work. The ELT cycle is not necessarily completed within a single engagement. CE and RO occur during engagements, while AC develops both during and after engagements. AE operates throughout the engagement at varying intensities, most concentrated in testing, where each adversarial assessment informs subsequent tests, and it continues between engagements as participants apply what they have learned to their own work. The framework treats engagements as connected cycles rather than self-contained learning events, with AE between engagements feeding into the next concrete experience.

This cyclical, between-engagement structure is not incidental. It stems from the nature of what is being learned. Because AI systems are nondeterministic and constantly evolving, conclusions from one engagement quickly become outdated. Model updates, prompt adjustments, or new adversarial techniques can render previous principles irrelevant before the next engagement. Learning is a moving target, with less emphasis on fixed findings and more on developing an adaptable capacity to build risk models in changing conditions. This perspective shapes the framework: engagements operate in ongoing cycles rather than as isolated events, and the shared reasoning that rebuilds these models serves as the lasting outcome.

AI purple teaming engagements unfold across three phases: scoping, testing, and reporting. Schön's (1983) reflection-in-action applies during the scoping and testing phases, as participants think through what to test, why, and how to adapt as the engagement unfolds. Reflection-on-action occurs during reporting, as participants look back on the engagement to draw insights and lessons learned. Scoping and reporting are the most collaborative phases, with participants engaged in shared reflection and decision-making. Testing is more asymmetric, with the security team engaged in CE and carrying it forward through artifacts and mental modeling. Across all phases, professional dialogue is the avenue through which reflective practice unfolds among participants who bring different expertise to a shared problem. Figure 1 illustrates how the four modes of the experiential learning cycle operate across these three engagement phases.

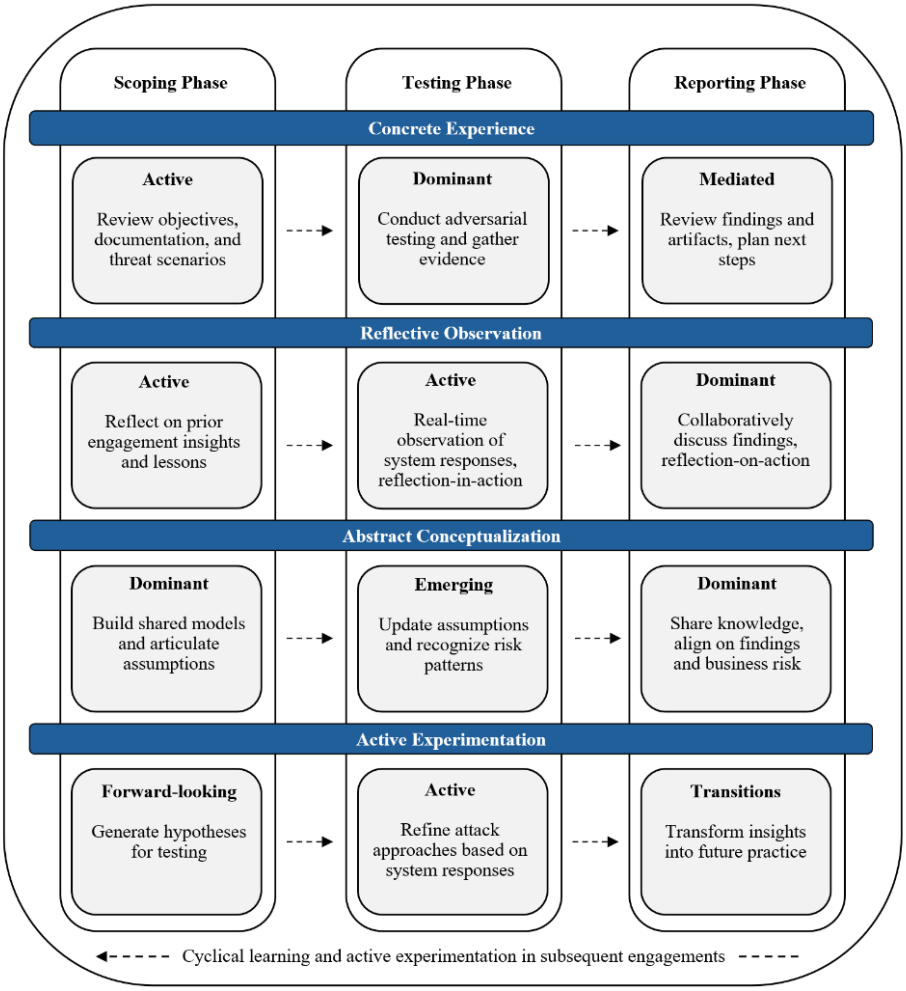


Figure 1. Experiential learning modes and intensities across AI purple team engagement phases

Scoping Phase

During the AI purple team scoping phase, security professionals, developers, and business partners collaborate to define what will be tested, identify the most critical risks, and determine which assumptions or problems to investigate. In this dialogue-heavy phase, the discussion focuses on architecture, trust boundaries, targets, threat models, prior incidents, acceptable risks, vulnerabilities, and standard rules of engagement. They discuss what is planned, why, and how. The team may also talk through testing methodologies, outline communication protocols for the engagement, and develop a timeline for the assessment to ensure comprehensive coverage and alignment with organizational goals.

In scoping, the ELT concrete experience (CE) mode typically involves participants reviewing system documentation, relevant vulnerabilities, prior incident data, threat scenarios, and perspectives from the business line, red team, and blue team. Reflective observation (RO) is collaborative and continuous; participants reflect together on prior engagements, known system weaknesses, and observations from similar systems. The third ELT mode, abstract conceptualization (AC), develops in real time as participants form shared principles and mental models about which threats matter and why. A developer discusses the foundation model, its parameters and configurations, system deployment, the tools the model can invoke, authentication mechanisms, and concerns such as threats, past incidents, and vulnerabilities. A business

partner might share what the system is used for in business terms, which outcomes the business considers most consequential, and what it deems an acceptable risk. A red teamer might ask about trust boundaries or potential threat scenarios, prompting developers to consider broader principles for where attackers may exploit and escalate privileges. A blue teamer may ask what types of logging and monitoring are in place, which fields are captured, where logs are stored, what retention applies, and which events are or are not logged. Blue teams need this to assess detection feasibility. A developer may realize for the first time that their application's permissions are broader than they had thought, or that they are missing important guardrails. In AC mode, transferable understanding occurs, in which participants not only learn from others' processes but also identify considerations they had not previously considered, such as a specific logging function, guardrail, risk, or threat scenario. The dialogue surfaces tacit knowledge (Nonaka, 1994) and builds AI risk awareness among all participants, producing an understanding of the system that no participant holds in full beforehand, and that is distributed across participants in a transactive memory system (Wegner, 1987). Active experimentation (AE) does not operate strongly within scoping itself; participants are not yet applying what they have learned in practice. The scoping phase produces hypotheses about where vulnerabilities will appear for testing, and AE will operate more strongly in subsequent work as participants apply what they learned to their own systems and processes.

Reflection-in-action operates throughout this phase, deepening the analysis in three ways that the ELT cycle alone does not capture. First, it highlights joint observations of system behaviors and real-time adjustments in reasoning as participants encounter new information, showing that scoping is an iterative thought process rather than a simple, sequential exchange of information. Second, it emphasizes professional dialogue as the key medium for shared reflection, with each participant's articulation prompting others to refine their thoughts. Third, it reveals the modeling dynamic as multidirectional rather than unidirectional from security to developers. As developers articulate design assumptions they had not previously made explicit, security professionals, in turn, share their threat reasoning. Reflection-on-action involves participants thinking back on previous engagements to inform current scoping and to share in the discussion.

The expertise developed during scoping is foundational. It produces a shared awareness of which threats matter, calibrated assumptions about how the system might respond under adversarial conditions, a mutual understanding of operational and organizational stakes, and a working set of mental models that participants will refine during testing.

Testing Phase

During the AI purple team testing phase, the red team conducts adversarial tests while the blue team monitors logs and alerts. AI red teaming techniques include prompt injection (where adversarial inputs manipulate the model's instructions and behavior), jailbreaks that bypass the model's safety measures, indirect prompt injection via retrieved or referenced external content, and trust boundary attacks, especially when models can use tools or access sensitive information. The blue team monitors the detection pipeline, prompt behaviors, and the development team's log-capture mechanisms. Developers join periodically to address questions about system behavior or clarify findings that require immediate attention, fostering cross-functional learning. However, developers typically do not observe testing in real time, and business partners are usually less involved at this stage.

Concrete experience (CE) is the dominant ELT mode during testing, with the red and blue teams directly engaging with the AI system's behavior in adversarial scenarios. The red team observes how the AI system responds to adversarial inputs and techniques, while the blue team assesses whether monitoring detects them. This phase generates raw experiential data through hands-on testing and analysis; however, it remains largely internal to security functions. Dialogue occurs within the security team and through periodic exchanges with the developers.

Artifacts produced during testing serve as the bridge between the security team's CE and the broader cross-functional reflection that occurs in reporting. These include findings, the rationale for assessment methods, screenshots of system behavior during attacks, and detection traces showing what monitoring captured or missed. Together, these elements form the foundation for collaborative reflection in the next stage. The framework treats this asymmetry as a feature rather than a limitation: it focuses CE within the security team, facilitating its integration into cross-functional reporting. Reflective observation (RO) occurs in real time as red and blue team members reflect on what just transpired, why an attack succeeded or failed, and what the system's responses indicate. It is also shared when findings are passed to the developers. AC takes shape as patterns and themes emerge from various tests. AE is inherent in the testing process. Each adversarial test serves as an experiment, with subsequent tests guided by earlier findings.

Reflection-in-action involves red teamers continuously adjusting test inputs based on system responses, while blue teamers constantly assess or challenge the effectiveness of monitoring. Schön's (1983) professional reasoning is highly visible during testing, initially primarily for the security team but later accessible to other participants through artifacts and reporting dialogue. The security team's active reflection, including observations, pattern recognition, and connections between behaviors, becomes retrospective analysis for others during reporting, facilitated by these artifacts. As the security team's CE accumulates, it serves as shared experiential material for the entire cross-functional group, contributing to the collective expertise that develops through continuous engagement cycles. Reflection-on-action is central: security professionals reflect on what testing revealed and organize and prioritize findings.

Testing develops expertise that is both internal to the security functions and latent within the cross-functional group. Within the security team, testing sharpens adversarial reasoning, validates or challenges detection coverage, and builds a practitioner-level understanding of the system's behavior under attack. For the broader group, testing yields experiential insights that inform collaborative discussion during reporting. The security expertise developed through testing is essential yet incomplete; it requires cross-functional integration during the reporting phase.

Reporting Phase

The reporting phase centers on the artifacts produced during testing and the reasoning behind them. It is also the engagement phase, during which the ELT cycle becomes fully visible across all preceding work as participants engage explicitly with the full arc of the engagement. The red team documents and prioritizes findings, rates risks, translates them into remediation actions, and presents them to the other AI purple team participants through collaborative dialogue. The blue team shares what they observed or did not observe during monitoring, and developers ask questions and provide operational insights. The engagement's outputs become collaborative learning material for all participants.

Concrete experience (CE) in reporting is mediated rather than direct. Security professionals and developers, often joined by business partners, engage with the testing experience through artifacts and reasoning rather than the live system itself, making concentrated CE accessible to the broader cross-functional group for the first time. What is shared in the report and discussion strongly shapes whether the engagement produces lasting change and follow-through. Reflective Observation (RO) is the primary ELT mode, characterized by intense collaboration. Security teams share testing observations through artifacts, and participants exchange information and reflect together on what was found, what it means, and next steps. A developer's experience learning how an attack succeeds against their system is different from simply being told that it could. Abstract conceptualization (AC) operates in close conjunction with RO, as participants establish shared principles for risk prioritization and remediation strategies through discussion. The dialogue is continuous, with the reporting meeting serving as a forum for ongoing discussion. Security experts explain

their reasoning for risk assessment and remediation, modeling security design principles that others then adopt and communicate within their teams. Active experimentation (AE) emerges as participants commit to corrective actions and share improvements to configurations, security controls, and monitoring. It transitions from a within-engagement to a between-engagement mode, generating CE for future cycles.

Reflection-on-action contributes to analysis in ways the ELT cycle alone does not. It foregrounds the retrospective and synthetic character of reflection in reporting, in which participants compare scoping assumptions with testing outcomes and translate that comparison into commitments to subsequent action. Reflection is a key mechanism that makes experiential learning actionable beyond the engagement itself.

The translational process is visible in the distinct roles participants play during reporting dialogue. Participants collaboratively review test observations, revisit assumptions established during scoping, and interpret the implications of these comparisons. Security professionals present findings and explain their prioritization logic. Developers contribute operational insights, including system constraints that contextualize the findings. Business partners articulate risk-acceptance reasoning and assess how technical results affect organizational risk. Modeling flows multidirectionally, as participants share and interpret one another’s reasoning across professional boundaries.

The reporting phase consolidates cross-functional expertise and completes the engagement cycle, making shared knowledge most visible and actionable. Participants collaborate in discussions and leave with a shared understanding of risk priorities, calibrated mental models of system behavior under adversarial conditions, and aligned remediation plans that address technical, operational, and business risks. The expertise built during this phase also supports the framework's cyclical nature: commitments made during reporting lead to experiments between engagements that create concrete experiences for future cycles.

Synthesis

In AI purple teaming, the ELT cycle operates collaboratively across cross-functional participants. Scoping enhances shared threat awareness through anticipatory reflection and a conceptual understanding. Testing delivers a focused, tangible experience and yields findings and artifacts for cross-functional reflection. Reporting consolidates expertise through collaborative reflection-on-action on the findings and artifacts. The phases are interconnected, each building on the last and feeding the next. Across each engagement, concrete experience accumulates, reflective observation deepens, and abstract conceptualization consolidates. Active experimentation between engagements generates new concrete experiences that build cumulative expertise over time, including the collaborative ability to model AI risk as conditions evolve.

For example, an illustrative composite engagement drawn from recurring patterns across various purple team exercises involved a team with two model developers, a business partner, red team testers, and a blue team detection analyst monitoring logs and telemetry. During scoping, the developers’ explanations of their goals, model designs, and trust boundaries led testers to create threat models and attack strategies. The detection analyst's questions about logging prompted the developers and business partner to review telemetry and guardrails, revealing previously unnoticed gaps. During testing, red team prompts uncovered unexpected behaviors. In reporting, the red team explained their reasoning, causing developers to identify a design assumption that shifted their understanding of the problem. This collaborative reflection fostered shared principles for input handling and monitoring, guiding subsequent developer work.

This framework treats AI purple teaming as an experiential learning practice that unfolds among collaborating team members, each contributing distinct perspectives. Drawing on Schön's (1983) work on reflective practice in organizational settings, the framework holds that this cross-functional cycle fosters a shared understanding of security that goes beyond what traditional policy, training, and awareness programs

provide on their own. This positions AI purple teaming as a collaborative expertise-development practice rather than solely a vulnerability-testing activity. In this respect, it extends Kolb and Kolb's (2017) account of collective experiential learning to a role-differentiated organizational practice. The cycle's modes are carried by different participants and completed collaboratively rather than by any single learner.

Discussion and Implications

AI purple teaming is an organizational practice in which security professionals, developers, and business partners observe real attacks and defenses within their systems. The framework presented here complements established AI risk management approaches, such as the NIST (2023) AI RMF. Specifically, scoping supports Map (risk identification), testing aids Measure (risk assessment), and reporting helps Manage (risk response), while Governance is integrated throughout the engagement lifecycle. While risk management frameworks establish processes and responsibilities for managing AI risk, they pay comparatively little attention to how the cross-functional expertise required to carry out these activities develops through collaborative practice. The framework presented here addresses this dimension by examining how organizations can cultivate expertise through participation in AI purple teaming engagements. The framework may also inform other collaborative AI risk management activities, including AI impact assessments, AI incident response drills, and AI deployment evaluations.

AI red teaming and AI purple teaming both produce vulnerability findings and remediation recommendations. What AI purple teaming adds is cross-functional collaboration that builds shared understanding and supports cyclical organizational learning, sustaining security and detection capabilities over time. While a traditional finding remains relevant until fixed, an AI finding can be invalidated before the next engagement as conditions change. What endures across cycles is the reasoning behind the findings and the capacity to continuously improve how the team models AI risk. Argyris and Schön (1996) call this deuterolearning, or learning to learn: as participants revisit how they scoped, tested, and reasoned in earlier engagements, they refine their understanding of the system and of how they learn about it. The framework proposes that learning to learn is the distinctive value of AI purple teaming in the rapidly evolving AI era.

The benefit of AI purple teaming emerges in specific moments: when the red team explains why a particular attack succeeded, the development team identifies where to update configurations, application logging mechanisms, and access controls, and the blue team identifies where detection and monitoring can be refined. These moments depend on broad cross-functional participation. This distinguishes AI purple teaming as a practice from traditional purple teaming and from isolated red or blue team exercises. The framework reframes AI security and safety as building cross-functional expertise rather than as work done in isolation by security teams. This has implications for how organizations structure AI roles, invest in skill development, and understand the relationships among security, development, and business teams.

The human element is increasingly vital as AI security testing adopts more automation and agentic tooling. While automated methods efficiently detect vulnerabilities and validate detection coverage, the cross-functional expertise the framework analyzes is built through human dialogue, reflection, and shared reasoning. Insights such as mental models, surfaced assumptions, and improved understanding arise from collaborative discussions during scoping and reporting, which current automated tools do not fully replicate. As AI systems and threats evolve faster than testing tools can be updated, the adaptive capacity of cross-functional human engagement remains a key strength of AI purple teaming.

AI Purple Teaming Design Principles

Eight design principles follow from the framework's analysis of how the experiential learning cycle operates among participants in AI purple team engagements. Each principle traces back to specific dynamics within the framework and is intended to support the development of cross-functional security expertise and organizational resilience.

1. **Psychological safety** – Cross-functional learning depends on participants willingly sharing knowledge, admitting uncertainty about their systems, and honestly reflecting on assumptions and outcomes. Edmondson's (1999) research on psychological safety in team learning demonstrates that these dialogue and reflection processes depend on participants feeling safe enough to take interpersonal risks. Therefore, engagements should be guided by the understanding that AI purple teaming is a collaborative learning activity rather than a performance assessment.
2. **Cross-functional participation in scoping** – Engagements should include security, development, and business team members as active contributors throughout scoping. Scoping is the phase in which shared mental models develop through iterative dialogue and abstract conceptualization, with each participant surfacing tacit knowledge for others. If non-security participants join only at the reporting phase, this foundational learning cannot occur.
3. **Collaborative reporting dialogue** – Reporting should be approached as a collaborative dialogue rather than as a unilateral delivery of findings. Reflection-on-action depends on participants actively engaging with the findings together, rather than on security professionals simply presenting while others passively listen.
4. **Deliberate reflection** – Engagements should incorporate deliberate reflection moments. Findings represent concrete experience, while reflection transforms those experiences into expertise. The framework suggests that skipping reflection and abstraction, by moving directly from findings to remediation, would diminish the development of security expertise.
5. **Repeatable engagements** – Engagements should be repeated, with each building on previous cycles. One-time engagements yield one cycle's worth of learning, whereas repeated engagements build cumulative security expertise. Repetition also helps keep mental models current.
6. **Collaborative assumption examination** – Team members should identify and examine assumptions, not just vulnerabilities. Building on Schön's (1983) reflection-in-action, expertise develops by questioning underlying assumptions rather than merely correcting errors within them.
7. **Artifact development** – Engagements should produce artifacts that guide participants' future active experimentation. Reports, recommendations, and reasoning traces serve as inputs to participants' work between engagements, helping them apply their knowledge and build new, concrete experience.
8. **Reasoning visibility** – Security professionals should make their reasoning visible, not just their findings and conclusions. Modeling and presenting reasoning enables other participants to observe and absorb it. Delivering only conclusions short-circuits the modeling process. This is particularly important for AI systems because specific findings can be invalidated by model updates, even as the underlying reasoning remains useful and transferable.

How these principles are enacted varies by organizational context. In a large, regulated enterprise, psychological safety must be cultivated across formal hierarchies and specialized roles, and reflection is typically scheduled and documented. In a smaller, AI-focused team, cross-functional participation is often inherent, but deliberate reflection must be safeguarded against the pace of rapid iteration.

Limitations

This study has several limitations that bound the claims of the presented framework. First, the study is conceptual rather than empirical. The framework is grounded in practitioner experience and a synthesis of existing literature, but it has not yet been tested through systematic research. The dynamics described across the engagement phases are theoretical propositions about how cross-functional learning occurs in AI purple team engagements rather than empirically established findings. Future research should empirically evaluate the framework through interviews, surveys, case studies, observational research, longitudinal analyses of repeated engagements, or comparative organizational studies to assess its applicability and effectiveness in practice.

Second, the framework is limited to collaborative engagement phases (scoping, testing, and reporting) and to learning among cross-functional participants. It does not analyze learning that occurs solely within security team activities, such as internal red team coordination, adversary emulation planning, or detection engineering work that does not involve cross-functional team members. Although these areas are important in AI purple teaming practice, they fall outside the framework's scope.

Third, the framework draws on experiential learning theory and reflective practice as its analytical foundation. Other theoretical lenses, such as communities of practice, could illuminate aspects of AI purple teaming that this framework does not emphasize. The focus on ELT and reflective practice is driven by their cyclical nature, which closely aligns with the engagement phases. Choosing different theories may yield alternative interpretations.

Fourth, the framework is based on the current state of AI purple teaming as of 2026. Because the field is evolving rapidly, the framework's specific applications may need to be adapted as AI systems and security practices continue to advance.

Conclusion

This paper presents a framework for AI purple teaming as a collaborative, experiential learning practice. Building on Kolb and Kolb's (2017) experiential learning theory and Schön's (1983) reflective practice, the framework examines how cross-functional security expertise develops through dialogue, observation, reflection, and conceptualization across the scoping, testing, and reporting phases of security testing. It reframes AI purple teaming as a practice for developing security expertise, in which security professionals, developers, and business partners build a shared understanding through collaborative engagement.

Eight design principles emerge from this analysis, offering practical guidance for organizations structuring AI purple team engagements. Together, these principles and the framework complement existing AI risk management approaches by embedding the development of organizational security expertise within the practices that cultivate it. As AI deployment accelerates and AI security and safety incidents rise, organizations increasingly require cross-functional security expertise to interpret evolving risks across technical and operational contexts. As reliance on automation and agentic tools in AI security testing grows, the importance of the cross-functional human collaboration outlined in this framework deepens.

References

Argyris, C., & Schön, D. A. (1996). *Organizational learning II: Theory, method, & practice.* Addison-Wesley.

Bullwinkel, B., Minnich, A., Chawla, S., Lopez, G., Pouliot, M., Maxwell, W., De Gruyter, J., Pratt, K., Qi, S., Chikanov, N., Lutz, R., Dheekonda, R. S. R., Jagdagdorj, B., Kim, E., Song, J., Hines, K., Jones, D., Severi, G., Lundeen, R., … Russinovich, M. (2025). *Lessons from red teaming 100 generative AI products*. arXiv. https://doi.org/10.48550/arXiv.2501.07238

Dewey, J. (1938). *Experience and education.* Simon and Schuster.

Dror, O. (2025). *Agentic purple teaming: A new strategic agentic AI security solution.* LASSO Security. https://www.lasso.security/blog/lasso-agentic-purple-teaming

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, *44*(2), 350–383. https://doi.org/10.2307/2666999

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Stanislav, F., … Clark, J. (2022). *Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned.* arXiv. https://doi.org/10.48550/arXiv.2209.07858

Gillespie, T., Shaw, R., Gray, M. L., & Suh, J. (2026). AI red-teaming is a sociotechnical problem: Beyond its utility in detecting AI vulnerabilities and bias, red-teaming raises important issues around values, labors, and harms. *Communications of the ACM*, *69*(2), 88–95. https://doi.org/10.1145/3731657

Kolb, D. A. (2015). *Experiential learning: Experience as the source of learning and development.* Pearson Education.

Kolb, A. Y., & Kolb, D. A. (2005). Learning styles and learning spaces: Enhancing experiential learning in higher education. *Academy of Management Learning & Education*, *4*(2), 193–212. https://doi.org/10.5465/amle.2005.17268566

Kolb, A. Y., & Kolb, D. A. (2017). Experiential learning theory as a guide for experiential educators in higher education. *Experiential Learning & Teaching in Higher Education, 1*(1), 38. https://doi.org/10.46787/elthe.v1i1.3362

Lewin, K. (1951). *Field theory in social science.* Harper & Row.

Lockheed Martin. (n.d.). *The cyber kill chain: A Lockheed Martin overview.* https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

Majumdar, S., Pendleton, B., & Gupta, A. (2025). Red teaming AI red teaming. *Proceedings of the 2025 Conference on Applied Machine Learning for Information Security (Proceedings of Machine Learning Research), 299*, 66–86. https://proceedings.mlr.press/v299/majumdar25a.html

Meta. (2023). *Introducing purple llama for safe and responsible AI development.* https://about.fb.com/news/2023/12/purple-llama-safe-responsible-ai-development/

MITRE Corporation. (n.d.-a). *MITRE ATLAS*. https://atlas.mitre.org/

MITRE Corporation. (n.d.-b). *MITRE ATT&CK*. https://attack.mitre.org/

Nairan, S., Rayavaram, P., Morales-Gonazel, C., Harper, M., Abbasalizadeh, M., & Vellamchety, K. (2025). Practical cybersecurity education: A course model using experiential learning theory. *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V (SIGCSE '25)*, 819–825. https://doi.org/10.1145/3641554.3701922

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://www.nist.gov/itl/ai-risk-management-framework

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(1), 14–37. http://www.jstor.org/stable/2635068

OWASP Foundation. (2024). *OWASP Top 10 for large language model applications.* https://owasp.org/www-project-top-10-for-large-language-model-applications/

OWASP Foundation. (2025). *OWASP Top 10 for agentic applications for 2026.* https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). Red teaming language models with language models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* 3419–3448. https://doi.org/10.18653/v1/2022.emnlp-main.225

Routin, D., Thoores, S., & Rossier, S. (2022). *Purple team strategies: Enhancing global security posture through uniting red and blue teams with adversary emulation.* Packt Publishing.

SCYTHE. (2026). *PTEFv4: Purple team exercise framework (PTEF).* https://scythe.io/ptef

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action.* Basic Books.

Sotiropoulos, J. (2024). *Adversarial AI attacks, mitigations, and defense strategies: A cybersecurity professional's guide to AI attacks, threat modeling, and securing AI with MLSecOps.* Packt Publishing.

Stanford HAI. (2026). *The 2026 AI index report.* Stanford University. https://hai.stanford.edu/ai-index/2026-ai-index-report

Van Buggenhout, E. (2024). Purple teaming: A comprehensive and collaborative approach to cyber security. *Cyber Security: A Peer-Reviewed Journal, 7*(3), 207–216. https://doi.org/10.69554/WGLN8212

Vandeleur, J. (2024). *Purple team exercise framework for continuous security validation.* SANS Institute. https://www.sans.org/blog/continuous-purple-teaming-practical-approach-strengthening-offensive-capabilities

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. R. Goethals (Eds.), *Theories of group behavior* (pp. 185–208). Springer-Verlag. https://doi.org/10.1007/978-1-4612-4634-3_9