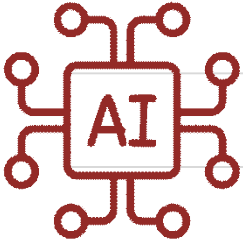
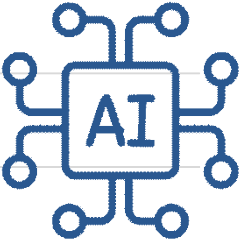


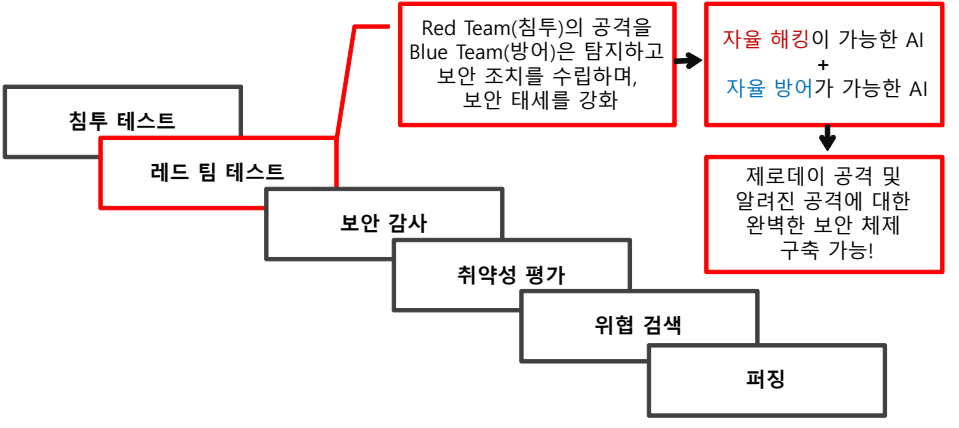
Purple Team AI Framework



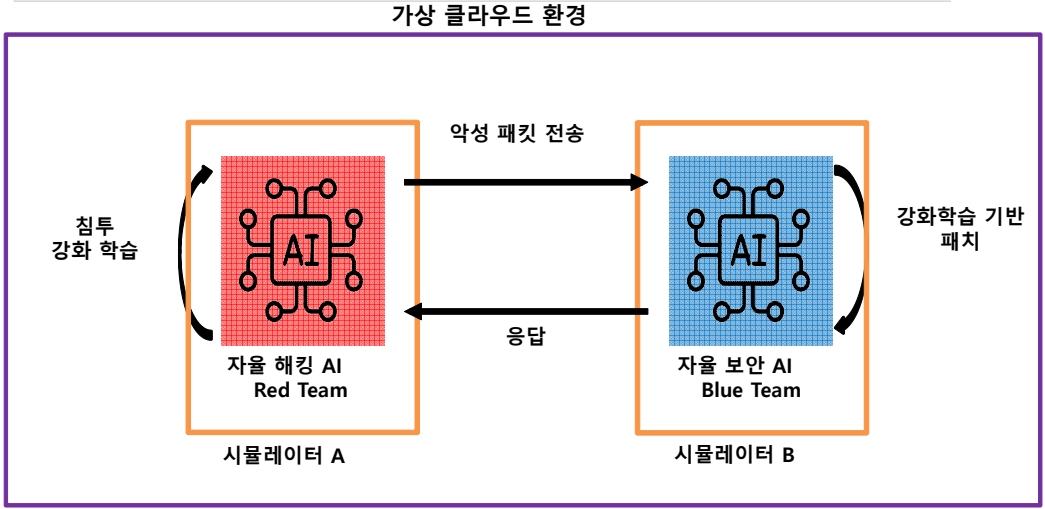
Hack My Brain
최민석
이정민

기획배경

현 보안 점검 시스템

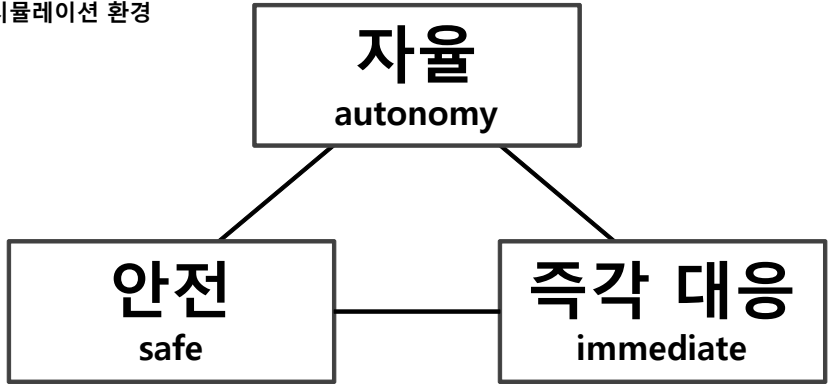


What is Purple Team AI Framework?



장점

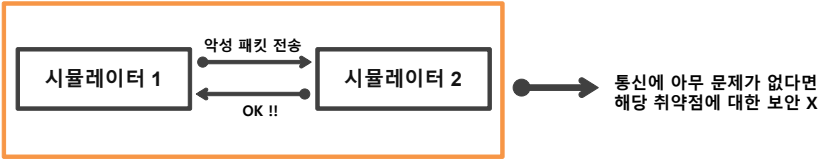
가상 시뮬레이션 환경



기획배경

BAS 기술 (Breach and Attack Simulation)

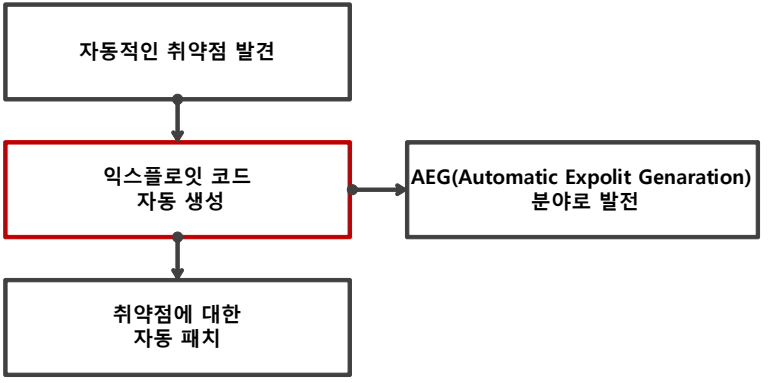
- **시나리오 기반의 자동화된 사이버 방어 시뮬레이션 테스트 도구**
- **ATT&CK 프레임워크 기반의 실제 발생한 사이버 공격 사례를 기반으로 보안 테스트를 해볼 수 있다.**
- **클라우드 및 가상 환경에서 시뮬레이션이 가능하기 때문에 기존 자산이나 서비스에 영향을 주지 않는다.**



BAS 시뮬레이터 동작원리

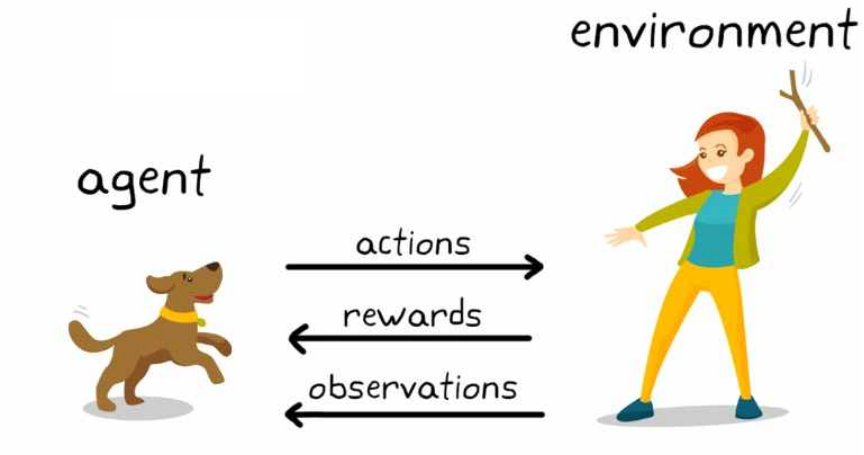
기획배경

자율 해킹 Red Team AI



구현 방법론

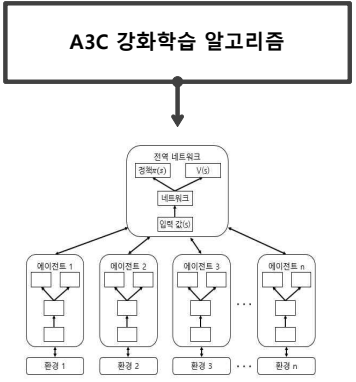
Red Team AI – 강화 학습

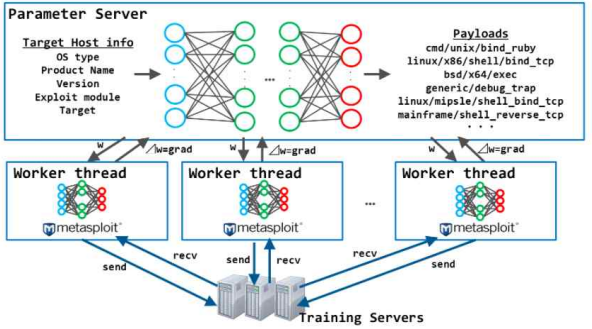


구현 방법론

Red Team AI – Exploit 코드 자동 생성

A3C 강화학습 알고리즘



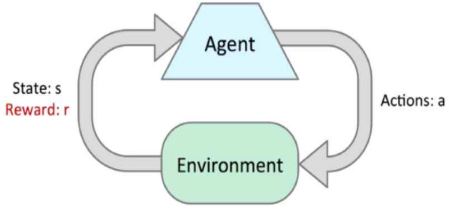


Advantage: A = Q(s, a) - V(s)
Advantage(A): 어떤 에이전트의 액션이 예상보다 얼마나 더 좋은 결과를 냈는지를 결정
V(s): 네트워크의 가치함수로 어떤 상태가 얼마나 좋은지를 나타낸다.
π(s): 일련의 액션의 확률 출력값 (정책)

구현 방법론

Red Team AI – Exploit 코드 자동 생성

Q- Learning



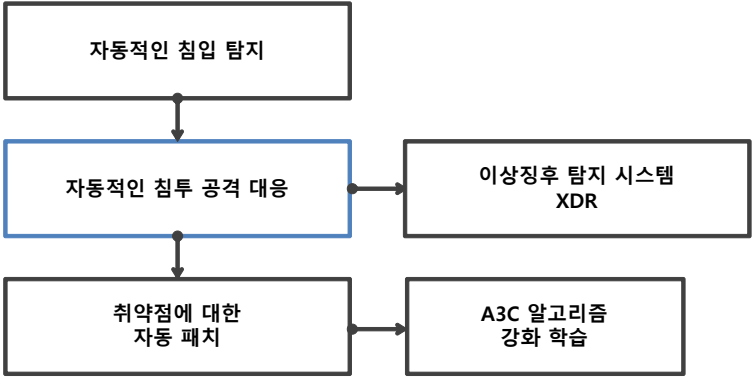
Model-Free Algorithm

- Agent가 Action을 통해 Expected sum of future reward를 최대로 하는 기능을 탐색

- 알고리즘은 Environment에 대해 알지 못하고,
Environment가 알려주는 Next State와 Next Reward를 수동적으로 획득

구현 방법론

Blue Team AI – 로직



구현 방법론

Blue Team AI – 자동적인 침입 탐지

자동적인 침입 탐지

AI기반 이상 징후 탐지 시스템

보안 이벤트 실시간 모니터링

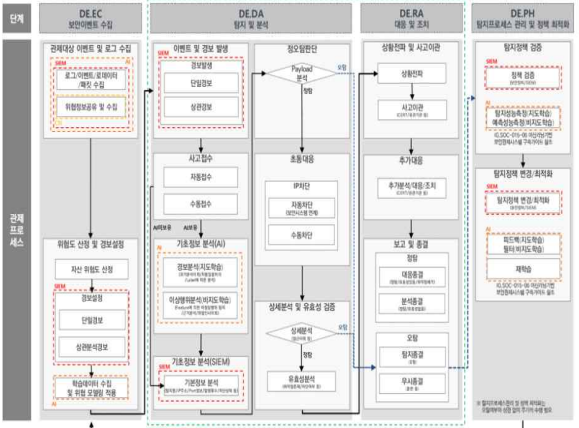
이벤트 발생

탐지 및 분석

대응

IP 차단(수동, 자동)
포트 차단

취약점 자동 패치



기대효과

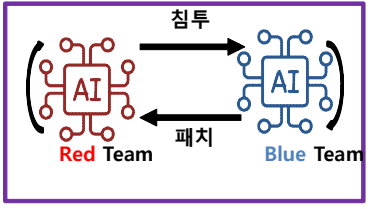
비용 절감

오탐 발생률 감소

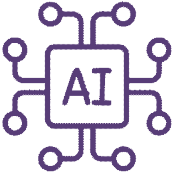
자산 보호

최종 모델

BAS기반의 강화학습 수행 환경



최종적인 AI 통합관제
Purple Team AI



• Purple Team
- Purple Team AI Framework를 통해 제로 데이 공격(알려지지않은 취약점) 대비

• Red Team
- 자율적인 취약점 탐색
- 발견한 취약점을 통해 Exploit
-> ZeroDay 공격 or 알려진 취약점

• Blue Team
- 자율적 보안침해행위 탐지
- 자동적 침해행위 대응
-> 보고서 및 자동 패치

참고 자료

[논문] 사이버 공격 시뮬레이션 기술 동향 – 이주영 외 2명
[논문] AEG: Automatic Exploit Generation – 차상길 외 2명
[논문] B2R2: Building an Efficient Front-End for Binary Analysis – kaist 차상길 교수
[논문] 인공지능을 활용한 네트워크 이상징후 탐지에 대한 연구 – 건국대 이국진 교수
[논문] Asynchronous Methods for Deep Reinforcement Learning
[GitHub] DefensiveOrigins/AtomicPurpleTeam
[GitHub] DeepSpaceHarbor/Awesome-AI-Security
[GitHub] 13o-bbr-bbq/machine_learning_security
[기사] Cybersecurity: Supervising Your AI With The Red Team
[기사] BAS, 동작원리 및 특장점..."예상 가능한 모든 위협, 사전에 체크하고 미리 차단 – 데일리시큐
[웹] MITRE ATT&CK Framework 이해하기 - 이글루시큐리티
[블로그] https://leemon.tistory.com/34