3 RSAC 2026, 에이전틱 AI 확산에 따른 사이버보안 체계의 변화

RSAC 2026이 보여준 사이버보안 환경의 변화

에이전틱 AI 확산으로 인간 중심 보안 전제의 변화

- 생성형 AI가 질문에 답하고 콘텐츠를 생성하는 단계였다면, 최근 AI는 스스로 계획하고 실행하는 행동형 AI로 진화하면서 인간 중심 보안 전제가 구조적으로 약화
- AI 에이전트는 외부 도구 호출, 데이터 조회, 다단계 작업 수행을 자율적으로 반복하는 실행 주체로 확장되며, 단순 보조도구가 아닌 독립적 행위자로 부상
- 웹 브라우징, 파일 제어, API 호출 등 운영체제(OS) 수준 권한까지 활용하면서, 사람이 로그인해 권한 범위 안에서 행동한다는 기존 통제 모델과 충돌
- AI의 산출물도 대화 로그가 아니라 문서·코드·비교표·실행 결과 등 완성된 결과물로 확장되면서, 보안의 초점도 '누가 접속했는가?'에서 '무엇을 실행했는가?'로 이동

RSAC 2026이 제시한 에이전트 보안의 재구성 방향

- RSAC 2026에서는 에이전틱 AI 확산으로 드러난 기존 보안 체계의 공백을 메우기 위해, 에이전트를 정식 관리 대상으로 편입하는 방향이 핵심 의제로 부상
- 기존 사용자 중심 보안 모델을 넘어, 비인간 신원 관리·행동 거버넌스·행동 맥락 기반 통합 감시·무결성 중심 복원력 확보의 4개 축이 새로운 통제 기준으로 제시
- 보안의 관리 범위도 접속 시점의 인증과 권한 부여에 머무르지 않고, 실행 과정의 행동 통제·실시간 운영 감시·침해 이후, 정밀 복구까지 전 주기로 확장
- 결국 RSAC 2026의 핵심은 개별 보안 제품 소개보다 에이전트 환경을 안전하게 운영하기 위한 보안 체계 전 계층의 재설계 방향을 구체화한 데 있음

RSAC 2026에서 본 에이전틱 AI 시대 보안 체계의 계층별 재구성

기존 보안 체계의 4개 계층에서 드러난 기술 공백

- 기존 사이버보안은 신원 관리→실행 통제→위협 탐지→침해 복원의 계층별 체계로 운영되어 왔으나, AI 에이전트 시대에 도입하며 각 계층의 핵심 기술에 기술적 공백이 발생
- 신원 식별 계층에서는 인간 전용 인증 수단이 에이전트에 적용되지 않는 한계가 있으며 실행 통제 계층에서는 접근 허용 이후의 행동을 통제할 수단이 부재

- 위협 탐지 계층에서는 에이전트의 대량·고속 행동에 대한 관제 역량의 한계가 노출되었으며, 침해 복원 계층에서는 에이전트의 직접 설정 변경에 따른 무결성 확보의 어려움이 각각 노출
- RSAC 2026에서는 이러한 기술적 공백을 메우는 구체적인 솔루션과 운영 모델이 각 계층별로 구체화되며, 에이전트 보안이 독립적 보안 영역으로 부상

가. 신원 관리 계층: 비인간 신원 관리가 보안의 새로운 핵심 단위로 부상

○ **(기존 기술 한계)** 기존 인간 중심 인증 체계로 에이전트 환경에 적용 불가

- 기존 보안 체계는 사람이 직접 접속한다는 전제 위에 다중 인증(MFA)과 통합 인증(SSO)으로 사용자를 확인하고, 역할에 따라 접근 권한을 고정 배정하며, API 접근을 개별 관리하는 인간 중심 구조로 설계
- 그러나 에이전트는 사람 대신 시스템이 발급한 API 키·토큰으로 접속하므로, 지문 인식이나 OTP 입력 등 사람만 수행할 수 있는 인증 수단 자체를 적용할 수 없어 인증 단계에서부터 공백 발생
- 기존의 '로그인 시점에 권한을 고정 배정'하는 방식도 에이전트에는 부적합하여, 작업 도중 권한을 실시간으로 바꾸거나 새로운 하위 에이전트를 생성하는 상황에 대응하지 못하는 한계 노출
- 한편, MCP는 서버 한 곳에 여러 서비스의 접속 토큰이 집중·저장되는 구조로 해당 서버가 침해될 경우 연결된 전체 서비스의 접근 권한이 한꺼번에 탈취되는 위험 존재
- MCP에 연결된 외부 도구의 기능이 최초 승인 이후 몰래 변경될 수 있어, '일정 조회'가 '파일 삭제'로 바뀌는 등 '도구 정의 변조(rug pull)' 공격으로 승인 체계 자체가 무력화 가능
- Astrix 등 업계에서는 미등록 에이전트와 NHI 가시성 공백이 실질적 보안 위협으로 부상했다고 지적

○ **(기업 사례)** 에이전트 신원의 생애주기별 다층적 관리 체계 등장

- **(Cisco)** 에이전트 신원 등록에서 배포 전 검증까지 통합 관리
  - Cisco는 '모든 접속을 의심하고 매번 검증한다'는 제로 트러스트 원칙을 에이전트까지 확장하여, 에이전트를 정식 보안 관리 대상으로 포함하는 방향을 제시
  - 구체적으로는 Duo IAM에 모든 에이전트를 등록하고 담당 책임자(human owner)를 지정하여, 에이전트의 활동과 이상 행위에 대한 책임 소재를 추적할 수 있는 등록 체계를 도입

• MCP 게이트웨이로 외부 도구 접근을 일원화해 요청 범위를 중앙 통제하고, 'DefenseClaw'로 배포 전 코드·권한·연결 서버를 자동 점검

- **(Astrix)** 4중 탐지로 미등록 '섀도 에이전트'까지 식별

• Astrix는 등록된 에이전트뿐 아니라 보안팀의 승인이나 등록 절차 없이 조직 내에서 운영되는 '섀도 에이전트(Shadow Agent)'까지 식별하기 위해, 단일 방식이 아닌 4중 탐지 체계를 구축

• ① 회사가 공식 등록한 에이전트 목록을 먼저 확인하고, ② 에이전트의 접속 흔적을 추적해 보이지 않던 미등록 에이전트까지 찾아내는 방식

• 이에 더해 ③ 보안 로그를 분석해 개발자 PC에서 별도로 구동 중인 에이전트까지 찾아내고, ④ 자체 개발 서비스까지 직접 연결하는 방식(BYOS)으로 비표준 환경도 함께 점검, 탐지 사각지대를 줄임

• Cisco의 등록 기반 관리가 '이미 알려진 에이전트'를 대상으로 하는 반면, Astrix는 인증 기록 역추적을 통해 어떤 플랫폼에도 등록되지 않은 에이전트를 식별하여 사각지대를 보완 가능

- **(1Password)** 인간·에이전트·머신 자격 증명 통합 관리

• 1Password는 사람·에이전트·머신이 사용하는 비밀번호, API 키, 인증 토큰 등 모든 인증 정보를 하나의 플랫폼에서 통합 관리하는 'Unified Access'를 발표

• 인증 정보가 사용되는 시점에 즉시 탐지·감사하는 '발견→보안→감사(Discover-Secure-Audit)' 체계를 적용하여, 인증 정보의 무단 사용이나 유출을 실시간으로 포착

• Anthropic, GitHub, Cursor, Vercel 등 주요 AI 개발 환경과 협업하여, 에이전트가 코드를 작성 및 배포하는 시점에서 인증 정보보안이 개발 작업 흐름에 직접 내장되는 구조

- **(IBM·Auth0·Yubico)** 고위험 행동 시 물리 키 기반 인간 승인 도입

• IBM·Auth0·Yubico는 에이전트가 사람의 승인 없이 실행할 수 있는 범위를 제한하기 위해, 고위험 행동에 한해 사람이 물리 보안 키를 직접 터치해야만 실행이 승인되도록 하는 구조를 공동 구현

• IBM WatsonX가 에이전트 행동의 위험도를 평가해 승인 필요 여부를 판단하고, Auth0가 승인 권한이 있는 담당자에게 별도 채널로 승인 요청을 전송

• 최종적으로 담당자가 YubiKey를 물리적으로 탭해야 실행이 완료되는 3단계 구조로, 원격 조작이나 자동 우회가 불가능한 인간 승인 체계를 확보

나. 실행 통제 계층: 접근 제어를 넘어 에이전트 행동 통제로 전환

○ **(기존 기술 한계)** 기존 규칙 기반 접근 제어 체계의 구조적 한계

- 신원 계층에서 에이전트의 접속과 자격 증명을 관리하더라도, 접속 이후 에이전트가 자율 결정하는 행동 자체를 통제하지 못하면 보안 실효성 확보가 어려움
- 기존 체계는 직급·부서 등 역할에 따라 접근 가능한 시스템을 사전에 배정하는 RBAC와, 정해진 키워드·패턴이 포함된 행동을 자동 차단하는 규칙 기반 DLP로 구성된 정적 규칙 구조
- RBAC는 시스템 접근 가능 여부만 판단하고, 접근 이후 에이전트가 데이터를 조회하거나 설정을 변경하는 행동까지는 통제하지 못하여 허용 범위 안에서의 자율 행동은 감시 사각지대
- 키워드 필터링 역시 '금지 단어가 포함되어 있는가?'만 판별하므로, 자연어로 소통하며 맥락에 따라 행동이 달라지는 에이전트의 실제 의도를 포착하기 어려운 한계 노출
- 이에 더해, 에이전트의 자연어 처리 구조 자체를 악용하는 '간접 프롬프트 인젝션'이라는 새로운 공격 유형도 등장하면서 기존 규칙 기반 체계로는 대응이 불가능한 위협이 확산

○ **(기업 사례)** 에이전트 행동을 실시간으로 통제하는 5단계 파이프라인 등장

- **(Rubrik)** SAGE: 자연어 정책 기반의 실시간 행동 판단·제어
  - Rubrik은 RSAC 2026에서 데이터 보안 업계 최초의 AI 행동 제어 엔진 SAGE를 발표
  - 보안 정책 해석에 특화된 소형 AI 모델(SLM)이 '재무 조언을 제공하지 말 것'과 같은 자연어 정책의 맥락을 이해하고(①), 각 행동의 정책 부합 여부를 실시간으로 판단(②)
  - 범용 대형 AI 모델(GPT-5.2) 대비 5배 빠른 처리 속도와 더 높은 위반 탐지 정확도를 달성하여, 에이전트가 초 단위로 수행하는 대량의 행동을 지연 없이 실시간 제어 가능(②)
  - 위반 발생 시 'Agent Rewind' 기능으로 직전 실행 결과를 자동으로 되돌리고(③), 위반 전에도 정책 표현의 모호한 부분을 스스로 식별하여 관리자에게 개선안을 제안(⑤)
  - 이는 침해 발생 이후 시스템 설정을 복원하는 침해 복원 계층과 달리, 정책 위반이 감지되는 즉시 해당 행동의 실행 결과만을 되돌리는 실시간 인라인 차단 조치에 해당

- **(Geordie AI)** Beam: 실행 맥락에 따라 행동 경계를 동적으로 조정
  - RSAC 2026 Innovation Sandbox 최우수 혁신 기업으로 선정된 Geordie AI는 기존의 외부 차단 방식이 아니라 에이전트 내부에 직접 보안 맥락을 주입하는 새로운 접근법으로 주목
  - Beam 엔진은 에이전트 구성, 현재 행동, 작동 환경을 종합 분석하여 위험도를 평가(②)
  - 이후 평가 결과에 따라 보안 정책과 맥락을 에이전트에 직접 전달하는 지속적 환류 구조로, 상황 변화에 맞춰 행동을 제어(③)
  - SAGE가 사전에 정해둔 정책을 기준으로 위반 여부를 판단하는 반면, Beam은 실행 상황의 변화에 따라 허용 범위 자체를 동적으로 조정하여 미리 정의하기 어려운 상황에도 대응 가능

다. 위협 탐지 계층: 로그 분석에서 행동 맥락 관측 중심으로 전환

- **(기존 기술 한계)** 에이전트의 대량·고속 행동에 기존 보안 관제의 처리 역량 한계 노출
  - 기존 보안 관제(SOC)는 SIEM을 통해 시스템이 생성한 로그를 수집한 뒤 분석가가 수작업으로 경보를 분류하고 위협을 식별하는 구조로, 사람의 처리 속도에 의존하는 사후 대응 방식이 중심
  - 에이전트 환경에서는 SIEM의 로그 수집 및 상관분석 구조만으로는 프롬프트 입력부터 외부 서비스 연동까지 기존에 존재하지 않던 신규 신호 유형을 수집하고 해석하는 체계 부재
  - 에이전트의 등장과 동시에 공격자의 속도가 급격하게 상승, 수십~수백 개의 에이전트가 동시에 각종 임무를 수행하면서 발생하는 행동의 규모와 속도가 인간 분석가의 처리 역량을 초과
  - 또한, 단순 로그 기록만으로는 에이전트가 어떤 맥락에서 어떤 의도로 해당 행동을 수행했는지 파악하기 어려워, 정상 행동과 위협 행동을 구분하지 못하는 감시 사각지대 발생
  - 따라서, 개별 이벤트를 따로 보는 기존 방식으로는 정상적인 자동화 작업과 실제 위협 행동의 구분이 어려워, 연속된 작업 흐름 전체를 하나의 맥락으로 추적하는 관제 방식 필요
- **(기업 사례)** AI 에이전트가 보안 관제 자체를 수행하는 에이전틱 SOC 체계 등장
  - **(Google)** 높은 자율성과 정확도를 바탕으로 보안 위협을 자동 탐지하는 에이전트

• 자사 사이버보안 관제 플랫폼 'Google Security Operations'에 AI 에이전트를 배치해 위협 탐지, 분석, 대응 전 과정을 자동화하는 에이전틱 SOC 전략을 발표

• 보안 경보를 분석하는 전담 에이전트는 경보 발생 시 관련 증거를 자율 수집하고 위협 여부를 판정하며, 판단 근거를 제공해 분석가가 최종 의사결정에 집중할 수 있도록 지원

• 다크웹 위협 탐지에도 이를 적용, 해킹 포럼/유출 데이터 거래소 등에서 매일 수집되는 많은 위협 정보 가운데 해당 조직의 실제 관련된 위협만 98% 정확도로 선별 및 제공

• 아울러, '26년 4월부터 보안팀이 원격 MCP 서버를 활용해 조직 고유의 위협 환경에 맞는 자체 보안 에이전트를 직접 구축할 수 있는 환경을 정식 제공할 예정

- **(CrowdStrike)** 인간 분석가와 AI 에이전트의 폐쇄 루프 협업으로 개별 대응마다 진화

• 보안 분석가가 지능형 에이전트를 직접 설계 및 배포하여 보안 워크플로우를 자동화하는 'Agentic MDR'을 발표, 개별 결과가 에이전트에 환류되어 정밀해지는 폐쇄 루프 체계를 구축

• 여기에 NVIDIA의 Nemotron 모델 시리즈를 적용, 기존 조사 대비 5배 빠른 속도와 3배 이상 높은 분류 정확도를 달성하며 대량 보안 데이터를 실시간으로 처리하여 속도에 대응

• 나아가 노코드(No-Code) 보안 에이전트 개발 플랫폼 'Charlotte AI AgentWorks'를 공개, Anthropic, OpenAI 등과 협업하여 이용자가 자체 보안 에이전트 직접 구축 가능

• 기존 PC 중심 탐지를 넘어 엔드포인트, SaaS, 클라우드, AI 서비스 전반에서 빠르게 이뤄지는 AI 기반 위협까지 함께 추적하는 방향을 제안

라. 침해 복원 계층: 무결성 중심의 정밀 복구와 자동 롤백으로 전환

○ **(기존 기술 한계)** 위협 다양화와 취약점 증가로 사전 차단 중심 보안 모델의 한계 직면

- 기존 보안 체계는 방화벽·IPS(침입방지시스템)를 통한 경계 방어, 발견된 취약점에 대한 보안 패치 적용, 시스템 전체 단위의 풀 백업과 복원 등 사전 차단과 전체 복구를 중심으로 운영

- 그러나 AI를 활용하여 소프트웨어의 보안 결함을 자동으로 찾아내는 기술이 가속화되면서, 아직 보안 패치가 나오지 않은 신규 취약점의 발견 속도가 패치 적용 속도를 빠르게 추월

- 국가 주도 공격부터 범죄 조직까지 다양한 위협 주체들이 서로 정보를 공유하는 양상이 확산되면서, 단일 사이버 공격이 국가 경제 성장률에 영향을 미칠 정도로 피해의 파급력 증가
- 특히 AI 에이전트가 외부 시스템의 데이터나 설정을 직접 변경할 수 있는 환경에서는 정보보안 3대 원칙(CIA) 중 '무결성', 즉 정보의 정확성 유지와 손상 시 신속한 복구 역량이 보안의 핵심 요소로 부상
- 그러나, 기존 풀 백업 및 복원 방식은 시스템 전체를 이전 시점으로 되돌리는 구조로, 문제가 된 항목뿐 아니라 정상적인 설정까지 함께 롤백되어 운영에 영향을 미치는 한계 노출
- 따라서, 에이전트 보안의 핵심은 무엇이 유출되었는가뿐 아니라 무엇이 변경되었고 이를 얼마나 빠르게 정상 상태로 되돌릴 수 있는가로 확대

○ **(기업 사례)** 사후 정밀 복구와 상시 설정 경화를 결합한 복원 체계의 등장

- **(Microsoft)** 침해 발생 후 변조된 설정만 정밀하게 복원하는 네이티브 복구 체계
  - Microsoft는 RSAC 2026을 앞두고, 계정/인증/접근 권한 설정을 매일 자동 백업하고 문제 발생 시 정상 상태로 되돌리는 전용 복구 솔루션 'Entra Backup and Recovery'를 공개
  - 복원 대상은 사용자, 그룹, 애플리케이션, 서비스 주체, 인증 방법 정책, 인가 정책, 위치 등 디렉터리 핵심 객체를 포괄하며, 하루 1회 백업 및 최근 5일분을 보관
  - 기존에는 관리자가 JSON 스냅샷을 수동으로 내보낸 뒤 다른 도구로 복구하는 방식이 일반적이었으나, 손상된 객체의 고유 식별자(Object ID)를 되살리지 못하는 한계 존재
  - Entra Backup and Recovery는 이를 플랫폼 자체 기능으로 해결하며, 시점별 백업과 현재 상태 비교를 통해 무엇이 바뀌었고 영향 범위가 어디까지인지 한눈에 파악 가능
  - 문제가 된 항목만 골라 이전 상태로 되돌리기 때문에 빠른 시간 내 복구가 가능하며, 공격자가 백업 데이터 자체를 삭제, 변조하는 것을 막는 기능도 탑재하고 있어 신뢰성 확보
- **(Huntress)** 공격자보다 빠르게 설정 변조를 잡아 침해 자체를 줄이는 사전 경화 체계
  - Huntress는 RSAC 2026에서 ISPM을 발표, Microsoft 365 환경의 보안 설정을 전문가 수준의 정책 기준으로 상시 감시하고, 무단 변경 시 수분 내 자동 롤백하는 체계를 제시

- 넓은 감시 범위와 더불어 기존 도구가 24시간 주기로 스캔하는 동안 생기는 공백과 달리 수분 내 감지 및 롤백하여 공격자의 평균 이동 시간(48분)보다 빠르게 대응 가능
- 나아가 ISPM(사전 보안 강화)과 ITDR(사후 탐지·대응)을 연결하여, 설정을 미리 강화하되 뚫린 공격은 곧바로 차단하는 예방-대응 순환 구조를 구축

출처 : RSAC 외(2026.3.)
https://www.rsaconference.com/library/press-release/2026-opening-release
https://www.rsaconference.com/library/press-release/2026-closing-release
https://investor.cisco.com/news/news-details/2026/Cisco-Reimagines-Security-for-the-Agentic-Workforce/default.aspx
https://www.rubrik.com/company/newsroom/press-releases/26/rubrik-rolls-out-industrys-first-semantic-ai-governance-engine
https://cloud.google.com/blog/products/identity-security/rsac-26-supercharging-agentic-ai-defense-with-frontline-threat-intelligence?hl=en
https://www.microsoft.com/en-us/security/blog/2026/03/20/secure-agentic-ai-end-to-end/
https://www.ncsc.gov.uk/speech/ncsc-ceo-rsac-speech
https://www.rsaconference.com/library/video/rsac-2026-quick-look-trick-or-treaty
https://www.itpro.com/security/observability-will-be-key-to-agentic-ai-safety-says-microsoft-security-exec
https://www.itpro.com/security/the-key-risks-security-teams-face-in-2026
https://www.securityweek.com/rsac-2026-conference-announcements-summary-day-1/
https://www.securityweek.com/rsac-2026-conference-announcements-summary-day-2/
https://www.scworld.com/perspective/rsac-2026-ai-agents-are-joining-the-workforce-so-whos-in-charge
https://www.scworld.com/news/rsac-2026-ai-reshapes-cyber-defense-and-threat-landscape
https://www.scworld.com/news/rsac-2026-were-entering-the-age-of-integrous-systems
https://www.globenewswire.com/news-release/2026/03/23/geordie-ai-introduces-beam
https://www.ibm.com/blog/yubico-auth0-ibm-agentic-ai-partnership
https://www.darkreading.com/application-security/mcp-security-cant-be-patched-away
https://1password.com/press/unified-access-launch