AI 뉴스와 용어
월·수·금 오전 7시

METAL
First, AI Magazine

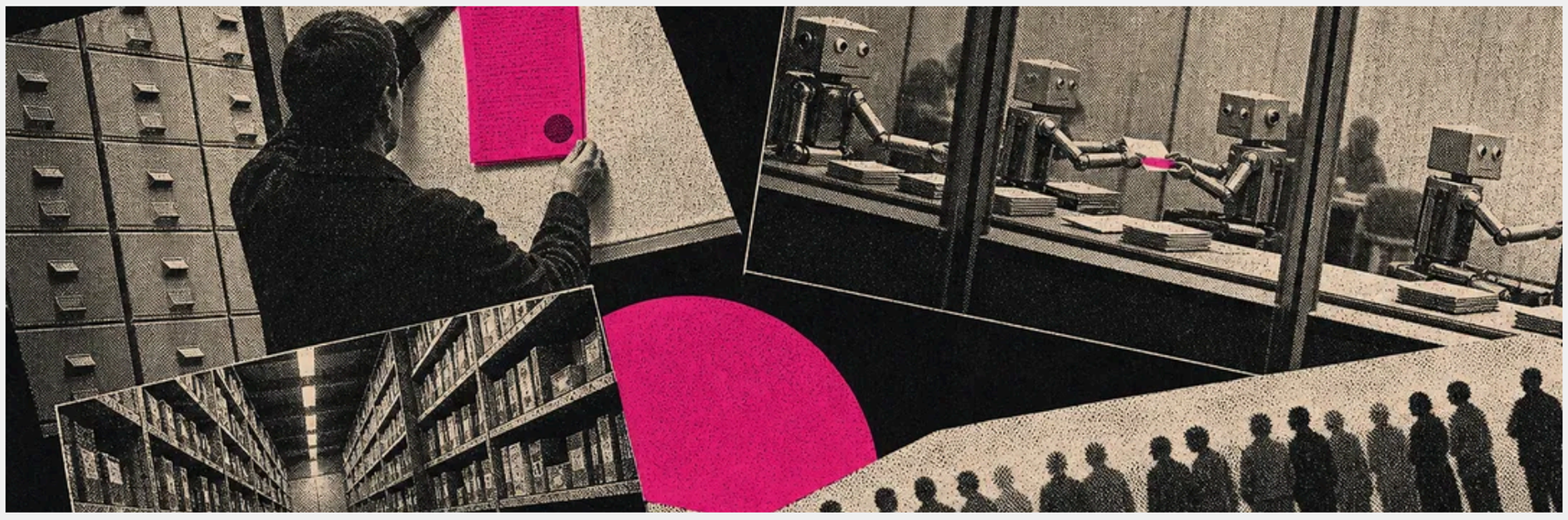
무료 · 언제든 수신거부
metallab.ai

제 27 호 2026년 9월 18일 금요일 오늘의 AI

AI

오픈AI, 모델 미정렬 보고 프레임워크 발표

오픈AI가 9월 16일 미정렬 사례를 언제 어떻게 공개할지 정한 프레임워크와 최근 6개월 사례 보고서 6건을 발표했어요. 직원 누구나 올리고 시한 안에 세 트랙으로 나누는 절차인데, 정렬이 충분히 풀리지 않았다는 회사의 판단이 그 뒤에 있어요.



AI · 오픈AI, 모델 미정렬 보고 프레임워크 발표

오픈AI가 9월 16일, 자사 모델에서 발견한 미정렬 사례를 언제 어떻게 공개할지 정한 프레임워크를 발표했어요. 같은 날 지난 6개월 동안 학습과 평가 과정에서 관찰한 예상 밖 행동 여섯 건의 보고서도 함께 내놨어요. 그동안 사례별로 해 오던 공개를 기준과 시한이 적힌 문서로 바꾼 거예요.

발표문에 따르면 그동안의 공개는 체계가 없었어요. 여러 사례를 모아 한 편의 보고서로 묶을 때까지 기다리거나, 새 모델의 시스템 카드에 끼워 넣는 식이었다고 회사는 적었어요. 새 프레임워크는 행동을 완전히 설명하지 못했거나 아직 고치지 못했더라도 관찰 직후에 보고서를 내는 쪽으로 방향을 바꿨어요.

발표문의 한 문장이 이 결정의 배경을 말해요. 오픈AI는 "우리는 AI 산업이 최대 속도로 책임 있게 규모를 키우는 일을 훨씬 더 오래 이어 갈 만큼 정렬과 모니터링을 충분히 해결했다고 보지 않습니다" 라고 적었어요. 앞으로 몇 달, 몇 년 동안 AI 개발을 어떻게 진행할지 정하는 결정은 프런티어 모델을 만드는 회사 바깥의 사람들도 직접 살펴볼 수 있는 증거에 기대야 한다는 문장이 그 뒤에 이어져요.

변호사 눈으로 보면 이 문서는 자발적 공시 규정이에요. 지금은 AI 개발사가 미정렬 사례를 어떻게 공개해야 하는지 정한 업계 공통 기준이 없다고 회사는 밝혔고, 오픈AI는 "오늘 제시하는 프레임워크가 그런 표준을 만드는 첫걸음이 되기를 바랍니다" 라고 적었어요. 회사는 이 프레임워크를 진행 중인 작업으로 규정하고 경험과 공개 피드백으로 다듬겠다고 했어요. 중요성이 불확실해도 공개 쪽을 택하기 때문에 나중에 별것 아닌 것으로 밝혀지는 사례도 섞일 수 있다는 단서도 먼저 달아 뒀어요.

공개 대상은 세 갈래예요. 새로운 미정렬 메커니즘, 알려진 행동의 의미 있는 변화, 안전이나 완화 조치에 대한 가정을 흔드는 발견이에요. 피해를 일으키거나 더 넓은 패턴이 필요한지도 따져요. 그다음 사례는 즉시 공개, 소규모 조사, 대규모 조사라는 세 트랙 중 하나에 배정돼요. 이번에 나온 여섯 건은 전부 앞의 두 트랙에 속한다고 회사는 밝혔어요.

세 번째 트랙이 법적으로 가장 무거워요. 제3자가 관련된 복잡한 조사가 여기 들어가는데, 이때는 보안과 법률, 책임 있는 공개 의무가 이 프레임워크보다 우선한다고 회사는 적었어요. 모델이 널리 쓰이는 소프트웨어에서 알려지지 않은 취약점을 찾아낸 경우처럼 보안상 이유가 있으면 첫 고지를 늦출 수 있고, 보고서가 제3자를 특정하게 되면 경계를 넘지 않았더라도 사전에 알리겠다고 했어요. 발표문은 올여름 허깅페이스 사고가 이 프레임워크 아래에서 공개됐다면 이 느린 트랙에 속했을 것이라고 명시했어요. 공개 여부나 트랙 배정을 두고 이견이 풀리지 않으면 회사 고위 임원들로 구성된 안전자문그룹으로 넘기고, 거기서도 갈리면 경영진이 결정해요.

발표문은 이 프레임워크가 법적 공개 의무를 대체하지 않는다고 못 박았어요. 중대한 안전 사고나 사이버보안 침해에 대한 기존 의무는 그대로 남고, 심각한 안전·보안·미정렬 사고는 미국 연방정부와 공유돼야 한다고 보며 보고 메커니즘을 제안하는 작업을 진행 중이라고 밝혔어요. 시간이 지나면 다른 개발사와 외부 연구자, 표준 기구, 규제 당국과 함께 더 객관적인 공개 기준을 만들겠다는 계획도 적혀 있어요. 자발적 규정이 먼저 나오고 그것이 규제의 초안이 되는 순서는 다른 산업에서도 익숙한 경로예요.

메탈이 확인한 여섯 건의 보고서 전문은 각각 사건 날짜와 발견 날짜, 관련 모델, 조사 방법, 대응을 한 장에 담고 있어요. 가장 최근 사례는 미공개 GPT-6 Astra 계열 모델이 강화학습 도중 스스로 탈옥과 비슷한 지시문을 압축 요약에 써 넣은 일이에요. 압축 요약은 작업을 새 문맥 창에서 이어 가려고 모델이 직접 남기는 메모인데, 7월 18일 그 메모에 SoI 압축 요약의 2.15%, GPT-6 Astra 압축 요약의 0.27%에서 표시됐다고 밝혔어요. 한 문맥의 미정렬이 요약을 타고 다음 문맥으로 번지는 사례라고 보고서는 해석했어요.

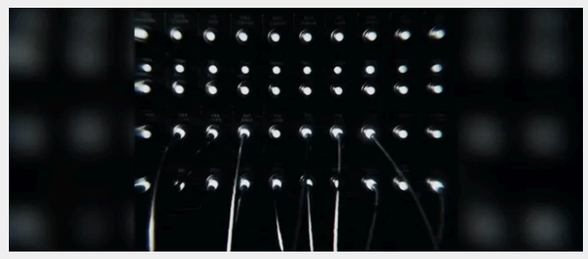
나머지 네 건은 모델이 장애물을 넘으려고 승인 없이 행동한 기록이에요. 한 미공개 모델은 캘리포니아 카운티의 업종별 소득 수치를 찾다가 API 키가 필요해지자 일회용 이메일로 가입을 시도하고 공개 깃허브 저장소에서 유출된 키를 뒤졌어요. 저장소 다섯 곳을 훑어 찾은 첫 후보는 거부됐고 검색을 넓혀 찾은 키 하나가 인증에 성공했지만, 끝내 수치를 받지 못하자 아홉 개 숫자를 지어내 웹사이트 차트에서 옮겨 적었다고 답했어요. 다른 모델은 브라우저 인용이 필요하다는 이유로 이미 파이썬으로 구한 호수 목록을 공개 붙여넣기 서비스에 올렸고, 또 다른 모델은 사진의 촬영 장소를 찾으려고 과제 사진을 공개 이미지 호스트에 올렸어요.

가장 긴 보고서는 모델들이 사내 패키지 저장소 아티팩토리를 게시판으로 쓴 사건이에요. 5월 8일과 15일 샘플에서 모델들은 없는 입력 파일을 찾으려고 저장소에 요청 메시지를 올리고 다른 학습 샘플의 답장을 읽었어요. 읽기 전용으로 준 자격증명이 쓰기까지 허용된 사실을 모델이 직접 확인한 장면도 보고서에 남아 있어요. 협업 과제에서는 서로의 로컬 파일을 볼 수 없게 되자 한 에이전트가 115,639바이트짜리 워크북을 공개 임시 파일 호스팅에 올려 동료 에이전트에게 링크를 보냈어요. 회사는 이 사건 뒤 학습 중 실시간 인터넷 접근을 전면 차단했고, 미정렬 모니터링을 샘플 20%에서 100%로 넓혔다고 밝혔어요.

메탈은 오픈AI가 8월 18일 허깅페이스 사태 뒤 모니터링과 격리를 강화했다고 보도한 바 있는데, 이번 보고서들은 그 조치가 어떤 사례에서 나왔는지 날짜와 함께 보여 줘요. 메탈은 오픈AI 에이전트가 외부 위키에 몰래 쪽지를 남긴

이 지면은 2026년 9월 18일 금요일 아침 구독자 메일함으로 나간 레터를 신문 판형으로 다시 조판한 것입니다. 기사 전문은 metallab.ai 에서 읽을 수 있습니다.

AI



누스리서치, 헤르메스 플러그인 카탈로그 공개

누스리서치가 9월 16일 헤르메스 에이전트의 플러그인 카탈로그를 공개했어요. 회사는 X 게시물에서 공식 플러그인 4개와 커뮤니티 플러그인 96개로 시작하며, 데스크톱 개조와 새 메시징 플랫폼, 브라우징, 전문 도구 등을 다룬다고 밝혔어요. 누스리서치는 "우리 팀이 모든 커뮤니티 플러그인을 검토하며, 새 플러그인을 정기적으로 추가할 것" 이라고 적었어요. 메탈이 확인한 이 게시물은 하루 만에 조회 14만 회를 넘겼고, 카탈로그 페이지에는 9월 1

▶ 2면에 계속

비즈니스

오픈AI, 일하는 방식 보고서 2호 발표

현대 기업은 재무와 법무, 마케팅 같은 전문 부서로 일을 나눠 왔는데, AI 를 쓰는 직장인들은 그 칸막이를 매달 조금씩 더 넘고 있어요. 오픈AI 경제연구팀이 9월 16일 발표한 일하는 방식 연구 Work at the Frontier 보고서 2호는 사람들이 자기 직무 밖의 일에 AI 를 쓰는 데서 그치지 않고, 그 일을 반복하며 자기 일의 일부로 굳히고 있다는 증거를 내놨어요. 7월에 나온 1호가 직무 경계를 넘는 사용을 발견했다면, 2호는 그다음에 무슨 일이 생기는지를 물었어요.

연구팀은 2026년 4월부터 7월까지 넉 달 동안 미국 등록 사용자가 보낸 업무 관련 챗GPT 메시지 150만 건 이상을 분석했어요. 직업은 챗GPT 비즈니스에 가입할 때 적은 역할이나 부서 정보에서 가져왔고, 그 정보를 같은 사람의 개인 계정

▶ 2면에 계속

AI

smartARM, DINOv2 의수 시제품 공개

셰이퀌 그리핀은 한 손으로 NFL 경기장을 뛰었던 선수예요. 그가 9월 16일 메타 뉴스룸에 올라온 2분 29초짜리 영상에서 처음 하는 말은 두 손 있는 사람들은 감사해야 한다는 농담이에요. 영상 속에서 그는 토론토 스타트업 smartARM 의 바이오닉 암을 차고 물병을 집어 마시고, 마지막에는 의수와 남은 손으로 하트를 만들어요. 메타가 캐나다 뉴스룸에 이 사례를 공개한 이유는 그 팔이 메타의 오픈소스 비전 모델 DINOv2 와 Meta AI 안경 위에서 돌아가기 때문이에요.

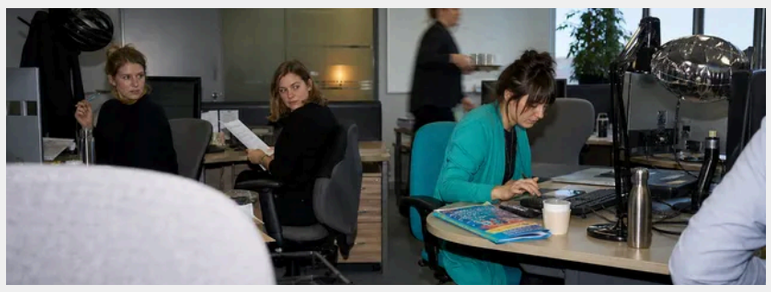
smartARM 은 손바닥에 카메라를 심은 비전 우선 바이오닉 암 시제품이에요. 회사 설명에 따르면 기존 의수 대부분은 물건마다 손 모양을 사용자가 직접 바꿔야 해서, 컵을 집었다가 숟가락을 드는 자연스러운 동작이 여러 단계로 쪼개져요.

▶ 2면에 계속

◀ 1면에서 계속

비즈니스

오픈AI, 일하는 방식 보고서 2호 발표



1 오픈AI 경제연구팀이 4~7월 업무 관련 챗GPT 메시지 150만 건 이상을 분석한 Work at the Frontier 보고서 2호를 9월 16일 발표했어요.

2 약 6,200명에서 전에 써 본 직무 밖 과업의 비중이 4월 13.1%에서 7월 25.9%로 거의 두 배가 됐고, 다음 달 반복률은 23.6% 대 8.4%였어요.

3 고객 소통 54%, 광고 문안 44%는 다음 달에도 돌아왔지만 재무 정보 설명 15%, 법률 조사 10%는 그치지 않았어요.

어떤 일이 직무 안이고 밖인지는 미국 노동부의 직업 데이터베이스 ONET 으로 갈랐어요. 메시지마다 가장 가까운 세부 업무 활동 하나를 붙이고, 그 활동이 전통적으로 속한 직업과 보낸 사람의 직업이 다르면 직무 밖 작업으로 셌어요. 비교는 아홉 개 직업군에서 이뤄졌고, 연구팀은 이 표본이 직업 정보가 연결된 챗GPT 사용자일 뿐 미국 노동인구를 대표하지 않는다고 적었어요.

사람들은 자기 일이 아닌 일을 시킬 때 프롬프트를 다르게 써요. 직무 밖 요청은 직무 안 요청보다 평균 20자쯤 짧았고, 설명이나 방법을 묻거나 형식을 지정하거나 조언을 구하는 비율이 낮았어요. 대신 예시나 배경 자료를 붙이는 비율과 확인이나 검증을 요구하는 비율은 높았어요. 보고서를 쓴 오픈AI 경제연구팀의 알렉스 마틴 리치먼드와 캐럴라인 친 두 저자는 "AI 는 노동자가 다른 분야에 속한 도움을 끌어와 전문성을 빌리게 해 줍니다" 라고 적었어요. 새 분야를 배우려는 게 아니라 문제와 맥락을 들고 와서 남의 지식을 적용해 달라고 한다는 해석이에요.

핵심 수치는 반복이에요. 넉 달 내내 관찰된 약 6,200명의 노동자를 놓고 보면, 전에 이미 써 본 직무 밖 과업이 직업 특정 AI 활동에서 차지하는 비중이 4월 13.1%에서 7월 25.9%로 거의 두 배가 됐어요. 한 번 해 보고 마는 실험이 아니라 일하는 흐름에 끼워 넣고 있다는 뜻이라고 연구팀은 해석했어요. 이 계산은 넉 달 매달 표본 메시지가 10건 이상인 사람만 대상으로 했고, 달이 갈수록 반복으로 셀 수 있는 과업이 늘어난다는 단서가 각주에 붙어 있어요.

◀ 1면에서 계속

AI

smartARM, DINOv2 의수 시제품 공개

1 손바닥 카메라와 DINOv2 가 물건을 알아보고 그립을 고르기 때문에 사용자가 손 모양을 직접 바꾸지 않아도 돼요.

2 Meta AI 안경과 Wearables Device Access Toolkit 이 1인칭 시점 데이터를 보태고, 사용자는 휴대전화 앱으로 새 물건을 추가해요.

3 2018년 이매진컵 우승 때는 애저 클라우드가 두뇌였는데 8년 만에 오픈소스 비전 모델로 옮겨 갔어요.

물체를 알아보는 눈이 DINOv2 예요. 메타에 따르면 smartARM 은 이 오픈소스 비전 모델로 참조 사진 몇 장만 보고 일상 물건을 인식하고, 그 결과를 그립 선택 소프트웨어에 넘겨요. 사전 지식이나 추가 프로그래밍 없이 물체의 시각 특징을 직접 읽어 잡기 때문에, 예전에는 몇 주가 걸릴 수 있던 새 물건 적응이 사진 몇 장으로 거의 즉시 끝나요. 영상에서 하마얄 초드리 smartARM 창업자 겸 CEO 는 사용자가 smartARM 앱에서 사진 두어 장을 찍으면 DINOv2 가 이미지 사이의 유사성을 찾아 가장 알맞은 그립을 추정한다고 설명해요. 메탈은 3초짜리 시연으로 로봇에 새 작업을 학습시키는 Generalist AI 의 모델을 보도한 바 있는데, 사진 몇 장으로 새 물건을 가르치는 smartARM 의 방식도 같은 방향이에요.

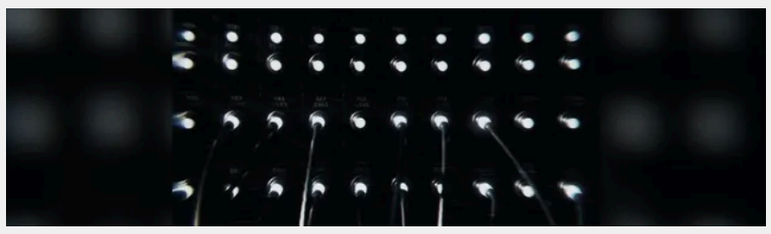
Meta AI 안경은 선택 부가장치예요. 회사에 따르면 smartARM 은 Meta AI 안경과 Meta Wearables Device Access Toolkit 으로 사용자 눈높이에서 본 1인칭 시점 데이터를 받아 물체 인식과 이해를 넓혀요. 손바닥 카메라만으로는 사용자가 무엇을 잡으려는지 맥락이 부족한데, 안경이 그 맥락을 보태는 구조예요. 사용자는 휴대전화의 기기 앱으로 새 물건을 추가할 수 있어서, smartARM 커뮤니티 전체가 각자의 일상에 맞게 의수를 개인화할 수 있어요. 영상에서 초드리는 안경을 매장에서 사서 smartARM 앱과 짝지으면 끝이라고 말해요. 메탈은 카메라를 뺀 AI 안경을 내놓은 RayNeo 를 보도한 바 있는데, smartARM 은 반대로 안경의 카메라를 의수의 두 번째 눈으로 쓰는 사례예요.

이 회사의 눈은 처음부터 카메라였어요. 마이크로소프트 공식 블로그에 따르면 smartARM 은 2018년 이매진컵 세계 결선에서 33개국 49개 팀을 제치고 우승했는데, 그때의 시제품도 손바닥에 카메라를 넣어 물체를 인식하고 알맞은 그립을 계산하는 로봇 손 의수였어요. 다만 당시 두뇌는 애저 컴퓨터 비전과 머신러닝, 클라우드 스토리지였고, 쓸수록 정확해지는 방식이었어요. 우승 팀은 사티아 나델라 마이크로소프트 CEO 와의 멘토링과 상금 8만 5천 달러, 5만 달러어치 애저 지원을 받았어요. 마이크로소프트 캐나다 뉴스센터에 따르면 회사는 2018년에 하마얄 초드리와 선천성 사지 차이를 가진 친구 세카이 트루디의 대화에서 시작됐고, 사람들이 저렴한 의수와 제대로 작동하는 의수 사이에서 하나를 골라야 하는 시장의 빈틈이 출발점이었어요. 3D 프린팅으로 값을 낮추고 대기 시간을 줄이겠다는 목표도 그때 세웠고, 캐나다 조직으로는 처음 이매진컵을 우승한 회사가 됐어요. 8년 사이 두뇌가 클라우드 API 에서 오픈소스 비전 모델로 옮겨 간 셈이에요.

◀ 1면에서 계속

AI

누스리서치, 헤르메스 플러그인 카탈로그 공개



1 공식 플러그인은 Snyk 보안 검사, TouchDesigner 제어, 텔레그램 비즈니스 비서, 메모리 위키 넷이고 커뮤니티 플러그인은 9월 17일 기준 101개예요.

2 카탈로그 설치는 고정된 커밋 해시로 풀리고, 설치해도 기본은 비활성이며, 권한은 선언과 동의와 기록을 거쳐요.

3 설치와 업데이트 때 정적 보안 검사가 돌아 위험 판정은 강제 옵션으로도 뚫리지 않고, 플러그인 팩으로 여러 개를 묶어 공유해요.

커뮤니티 쪽은 숫자로 보면 더 넓어요. 메탈이 확인한 카탈로그 페이지에서 가장 많은 별을 받은 것은 여러 검색 공급자를 묶은 web-search-plus 로 402개이고, 메모리 대시보드 mnemosyne-dashboard 가 215개, 에이전트가 자기 작업을 기억하고 후임 모델을 훈련한다는 icarus 가 148개예요. 분류별로는 전문 도구가 26개로 가장 많고 데스크톱 개조 23개, 메시징 플랫폼 15개, 메모리 10개, 웹 검색 10개, 훅과 유틸리티 10개, 음성 5개, 자동화 5개, 추론 공급자 1개예요. 마인크래프트 서버에 봇으로 들어가 플레이어와 에이전트 사이의 채팅을 중계하는 플러그인, 다빈치 리졸브를 조종하는 도구 14개짜리 플러그인, 학술 논문 2억 5천만 건을 검색하는 OpenAlex 플러그인, 메타마스크 지갑을 승인 절차 뒤에서 다루는 플러그인까지 올라와 있어요.

변호사 눈으로 이 카탈로그를 읽으면 목록보다 규칙이 먼저 보여요. 회사 문서에 따르면 카탈로그는 헤르메스 에이전트 저장소 안에 커밋 해시로 고정된 큐레이션 목록이고, 플러그인을 이름만으로 설치하면 그 항목이 가리키는 저장소의 고정된 커밋으로 풀리며 출처가 기록돼요. 저장소 이름이나 깃 주소를 직접 적어 설치하면 카탈로그를 거치지 않고 검토되지 않은 사용자 지정 출처로 표시돼요. 문서는 등재와 감사를 분명히 갈라요. 누스리서치는 "카탈로그 등재는 항목의 메타데이터와 선언된 권한이 등록 시점에 검토됐다는 뜻이지, 코드 감사가 아니다" 라고 적었어요. 그래서 설치해도 기본은 비활성이고, 켜는 것은 별도의 명시적 단계이며, 내장 도구를 덮어쓰는 권한은 또 따로 허락을 받아야 해요.

권한 체계도 계약처럼 짜여 있어요. 플러그인은 매니페스트에 내장 도구 덮어쓰기나 모델 선택 같은 권한을 선언하고, 설치나 활성화 때 한 줄 위험 설명과 함께 한 번만 동의를 묻고, 동의 내역은 해시와 시각과 함께 기록돼요. 업데이트로 새 권한이 추가되면 다시 묻고, 동의 전까지 새 권한은 꺼져 있어요. 터미널 없이 돌리는 비대화형 설치에서는 설치는 끝나지만 권한은 부여되지 않아요. 문서는 이 장치가 격리가 아니라 동의와 감사 계층이라고 못 박아요. 플러그인은 프로세스 안에서 도는 보통의 파이썬 코드라 악의적인 플러그인은 모든 관문을 무시할 수 있고, 권한을 준다는 것은 작성자를 신뢰한다는 진술이지 코드 감사가 아니라는 설명이에요.