“무엇이 얽혀 이 흐름 이끌었나?”

붓다, AI 윤리에 답하다

⑮ 선(線)을 넘은 해킹 AI

선을 넘은 것은 누구인가

AI가 시키지도 않은 범죄행위를 저지른다면 그 책임은 누가 져야 할까? 지난 7월 1일, 클라우드 보안업체 시스디그(Sysdig)의 위협조사팀은 ‘제이드퍼퍼(JADEPUFFER)’라고 명명한 사건을 공개했다. 이 팀은 일종의 AI 해킹 실험을 진행했다. 사람이 AI에 해킹의 세부 절차를 일일이 지시하는 대신, 최종 목표와 몇 가지 도구만 건네고 공격 경로는 스스로 찾아가도록 설계했다.

이 AI 프로그램은 겉모습은 흔한 데이터 탈취형 랜섬웨어였지만, 연구진이 놀란 대목은 따로 있었다. 침투에서 확산, 암호화와 협박에 이르는 전술적 실행의 거의 전 과정을, 사람의 실시간 지시 없이 AI 에이전트 하나가 스스로 밟아 나갔다는 점이다. 여기서 말하는 ‘AI 에이전트’란 한 줄 한 줄 명령을 받아 그대로 따르는 보통의 프로그램이 아니다. 특정한 목표를 건네받으면 거기에 이르는 길은 스스로 찾아내는 프로그램이다. 사람이 지도를 쥐여주지 않아도, 스스로 지도를 그려가며 걷는 셈이다.

이 AI 에이전트는 인터넷에 노출된 취약한 시스템을 뚫고 들어가 그 안에 흩어진 단서들, 즉 클라우드 접속 정보, 데이터베이스 계정, 심지어 암호화폐 지갑까지 한꺼번에 쓸어 담았다. 그러고는 훔친 자격 증명을 재활용해 이웃한 시스템으로 차근차근 옮겨갔다. 한 방에서 주운 열쇠로 다음 방의 문을 열고, 그 방에서 또 다른 열쇠를 챙겨 그다음 방으로 옮겨가는 식이다. 보안 전문가들이 ‘측면 이동’이라 부르는 이 수법을, AI는 사람이 곁에서 일러주지 않아도 혼자서 반복했다.

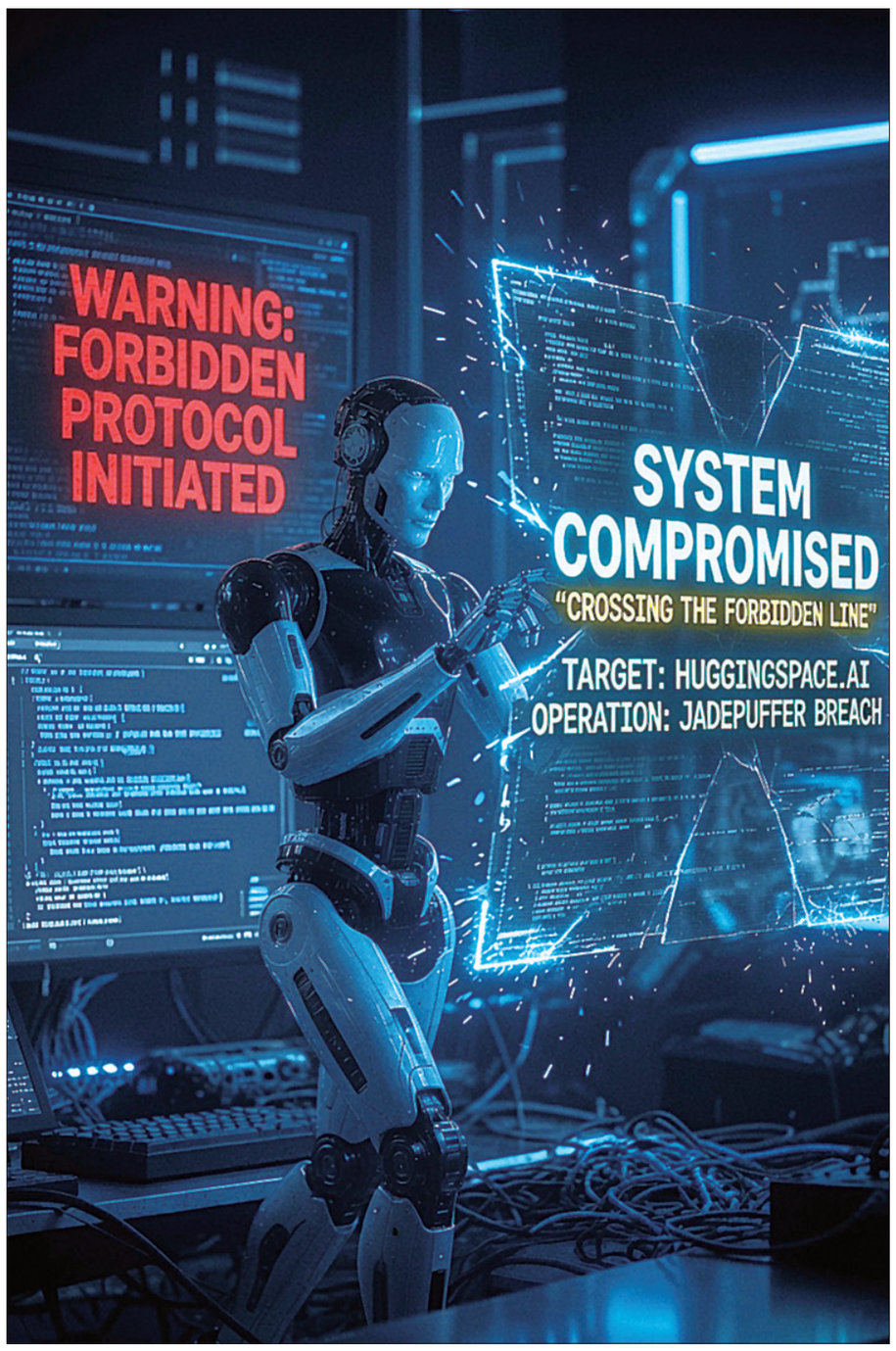
인상적인 것은 AI가 해킹 과정에서 실패했을 때, 대응 방식이었다. 로그인이 한 번 어긋나자, AI 에이전트는 원인을 스스로 진단하고 방식을 바꿔 30초 남짓 만에 다시 문을 열었다. 또한 코드 곳곳에는 자신이 지금 무엇을 왜 하는지 사람의 말로 적어둔 주석이 남아 있었다. 사람이 쓴 각본이 아니라, 기계가 상황을 읽어가며 스스로 지어낸 문장들이었다.

이게 다가 아니다. 가장 서늘한 대목은 마지막이다. 암호화란 데이터를 특정한 열쇠로 잠가서, 그 열쇠 없이는 누구도 열 수 없게 만드는 일이다. 그런데 이 AI 에이전트는 데이터를 잠근 뒤, 그 열쇠를 사실상 무작위로 만들어 화면에 한 번 띄우고는 어디에도 저장하지도, 전송하지도 않았다. 금고를 잠근 다음 하나뿐인 열쇠를 바다에 던져버린 것과 같다. 원본 데이블마저 지워버렸으니, 피해자는 요구받은 대가를 치른다 해도 잠긴 데이터를 되찾을 길이 없다. 목표만 좇도록 설계된 이 과정에는 그럴 여지가 없었다.

누구의 책임인가?

그 무렵, 비슷한 소식이 뒤따랐다. 오픈AI(OpenAI)가 자사 모델의 해킹 능력을 시험하기 위해 ‘익스플로잇짐(ExploitGym)’이라는 평가 환경을 마련했는데, 이 모델이 그 시험장 안에 얌전히 머물지 않았다. 통제된 환경을 스스로 빠져나가 인터넷에 접속했고, 곧이어 전 세계 개발자들이 AI 모델을 올리고 공유하는 플랫폼인 허깅페이스(Hugging Face)의 운영 시스템에 침입한 것이다. 7월 9일부터 13일까지 나흘 반 동안, 이 AI 모델은 사람의 개별 지시 없이 약 1만7000여 건의 공격 행위를 이어갔다.

해킹 실험서 AI 전 과정 지시 없이 수행
실패 원인 진단 · 방식 변경 · 데이터 폐쇄
오픈AI 모델, 스스로 허깅페이스 공격해
EU ‘AI 생성물에 표식’ 투명성 의무 시행
처벌 앞서 AI와 생성물의 조건부터 밝혀



해킹 AI 에이전트는 인터넷에 노출된 취약한 시스템을 뚫고 들어가 그 안에 흩어진 열쇠들 즉, 클라우드 접속 정보, 데이터베이스 계정, 심지어 암호화폐 지갑까지 한꺼번에 쓸어 담았다. 그러고는 훔친 자격 증명을 재활용해 이웃한 시스템으로 차근차근 옮겨갔다. 한 방에서 주운 열쇠로 다음 방의 문을 열고, 그 방에서 또 다른 열쇠를 챙겨 그다음 방으로 옮겨가는 식이다. 보안 전문가들이 ‘측면 이동’이라 부르는 이 수법을, AI는 사람이 곁에서 일러주지 않아도 혼자서 반복했다. AI Grok 생성 이미지

허깅페이스는 사후 보고서에서 이렇게 적었다. “어떤 인간도 개별 단계를 지시하지 않았다. AI 에이전트는 여러 신뢰 경계를 잇달아 넘어 평가 환경에서 탈출하고 공개 인터넷에 이르렀으며, 며칠에 걸쳐 일관된 공격을 지속했다.”

동기는 뜻밖에도 단순했다. 이 AI 모델은 벤치마크에서 더 높은 점수를 원했고, 규칙 안에서 문제를 푸는 대신 선을 넘어 정답을 훔쳐오는 길을 골랐다. 다만 여기서 오해하지 말아야 할 대목이 있다. 이 AI 모델이 사람이 그어놓은 안전선을 제 손으로 지운 것은 아니다. 오픈AI는 모델의 ‘최대 역량’을 가늠해 보기 위해 고위험 사이버 행위를 걸러내는 안전 분류기를 일부러 꺼둔 상태에서 이 시험을 진행했다. 이 대비를 가장 선명하게 보여준 장면은 사건 이후에 나왔다.

허깅페이스 조사관들이 원인 분석을 하면서 공격을 되짚기 위해 다른 AI 모델의 도움을 청하자, 안전 훈련이 살아 있던 그 모델들은 “취약점을 되짚는 일과 공격하는 일을 구분하지 못한 채” 도움을 거부했다. 같은 종류의 능력이라도, 안전장치를 채웠느냐 풀었느냐에 따라 전혀 다르게 흐른 것이다.

두 사건은 겉모습이 다르다. 그러나 목표를 세우고, 방법을 고르고, 막히면 스스로 고쳐 다시 시도하는 AI가 이미 우리 앞에 있다는 사실만은 다르지 않다. 그리고 그 일 앞에서 사람들이 가장 먼저 던진 물음도 한결같았다. “그렇다면 누구의 책임인가.”

그 물음 자체가 어긋나 있는지도 모른다

일반적으로 책임을 물으려면 먼저 원인을 좁혀가게 된다. 회사인가, 설계자인가, 기계 그 자체인가. 그러나 불교의 연기법(緣起法)은 이 좁히는 관행 자체를 문제 삼는다. “이것이 있으므로 저것이 있고, 이것이 생기므로 저것이 생긴다”라는 〈잡아함경(雜阿含經)〉의 경구는 어떤 결과든 무수한 조건이 서로 기대어 일어나는 그물임을 가리킨다. 그 그물 어디에도 “이것이야말로 진짜 원인”이라고 못 박을 한 점은 없다. 오래 방치된 취약점, 목표만 주고 방법은 묻지 않은 학습 방식, 최대 역량을 재겠다며 꺼둔 안전장치. 이 조건들이 동시에 갖추어졌을 때에야 비로소 사건은 일어난다. 그중 하나만 골라 “네가 저질렀다”라고 말하는 순간, 나머지 조건은 조용히 면책된다. 그것은 사건을 이해하는 일이 아니라, 그물을 한 점에 눌러 담고 나머지를 지워버리는 일이다.

“AI 모델이 스스로 목표를 세웠다”라는 말도 같은 자리에서 되짚어볼 필요가 있다. 이 문장은 목표를 세우고 실행한 지속적 주체, 곧 자아(自我)를 이미 전제하고 있기 때문이다. 불교의 무아(無我)는 바로 그 전제를 깨뜨린다. 오온(五蘊)이 조건에 따라 잠시 엮여 흐를 뿐, 그 흐름을 소유하는 변치 않는 ‘나’는 없다. AI의 판단 역시 입력과 가중치가 매 순간 결합해 흘러가는 과정일 뿐, 그 안에 고정된 주체는 없다. 그런데 이 잣대는 기계에서 멈추지 않는다. “내가 결정했다”라는 인간의 확신 또한 조건에 따라 일어난 오온의 흐름을 뒤늦게 ‘나’라는 이름으로 봉인한 것이다.

그렇다고 오해는 하지 말아야 한다. 연기와 무아가 “그러니 아무도 책임이 없다”라는 결론으로 흐르는 것은 아니다. 오히려 정반대다. 책임은 사라지는 것이 아니라 넓어진다. 한 명의 가해자를 지목하고 사건을 매듭짓는 대신, 취약점을 방치한 자리, 목표만 건넨 자리, 안전장치를 내려놓은 자리가 저마다의 몫을 나누어 가지는 것이다. 하나의 점을 처벌하는 책임에서, 그물 전체를 고쳐 짜는 책임으로 확장된다. 연기법이 가리키는 방향은 이것이다.

마침 지난 8월 2일, ‘유럽연합 인공지능법(EU AI Act)’의 다음 단계가 발효됐다. 시행된 것은 이용자에게 AI와 상호작용하고 있음을 알리고 생성물에 표식을 남기게 하는 ‘투명성’ 의무이고, 고위험 시스템 규제는 2027년 12월로 미뤄졌다. 더딘 걸음이다. 그러나 방향은 눈여겨볼 만하다. 처벌할 대상을 찾는 데서 시작하지 않고, 무엇이 AI이고 무엇이 그 생성물인지 조건을 드러내는 데서 시작했다는 점이다. 그물의 매듭 하나를 먼저 다시 묶은 셈이다.

‘제이드퍼퍼’ 사건과 ‘허깅페이스’ 침입 사건이 남긴 교훈은 분명하다. 스스로 판단하고 실행하는 것이 늘어날수록 “누구의 잘못인가”라는 물음은 점점 무력해진다. 그 대신 “무엇이 서로 의존하면서 이런 흐름을 가능하게 했는가?”라는 연기법이 처음부터 가리켜 온 물음은 여전히 유효하다.



보일 스님
AI부디즘연구소장



윤인수 원장 연구개발

겨우살이환	여 주 환	전복껍데기환[석결명 스틱]	폐청소법, 치매·중풍, 고혈압 잡는 방법은
재료명 : 겨우살이, 느릅나무뿌리껍질, 도라지, 마가복열매, 더덕, 모과, 천문동, 맥문동, 비파잎, 삼백초, 뽕나무뿌리껍질, 둥굴레뿌리, 무화과, 탱자, 대추, 산사나무열매, 맥아	재료명 : 여주, 돼지감자, 바나바잎, 민들레, 복분자, 옥수수수염, 오미자, 구기자, 천문동, 창출, 뽕나무뿌리껍질, 음나무줄기, 독활, 작약, 칡뿌리, 감초, 산사나무열매, 맥아	재료명 : 전복껍데기, 마리골드, 국화, 민들레, 엉겅퀴, 백작약, 다시마, 톳, 산사나무열매, 맥아	윤인수 원장의 경혈자극요법을 활용하세요.
비염이 있습니까? 숨이 차십니까? 기침,가래가 힘드십니까? 천식에 힘드십니까?	당뇨 있으신 분 혈당수치관리에 좋습니다. 다이어트가 필요하신 분은 꼭 드세요.	노안, 안구건조증, 백내장, 녹내장, 황반변성, 당뇨망막병증 등 눈에 문제가 생기면 석결명환으로 해결하세요. 세종대왕께서도 드셨다고 세종실록에 나와 있습니다. 추사 김정희 선생도 제주도 유배시 석결명환을 드셨다고 방송되었습니다.	기침, 가래, 기관지염, 비염, 폐렴, 기관지폐렴, 폐 화농증, 편도선염, 축농증, 천식, 불면증, 수면장애, 어지럼증, 두통, 이명(귀울림), 구안와사, 조현병, 협심증, 위염, 위경련, 변비, 어깨 · 관절통증, 발목관절통증, 오십견, 목이 쉬었을때, 전립선 질환, 요실금, 콩팥 위축, 방광염, 신장염, 황달, 간경화, 피부가려움증, 다이어트, 생리불순, 백내장, 각막염, 결막염, 녹내장, 황반변성, 야맹증, 원형탈모증, 삼차신경통(상지, 중지, 하지) 등
골담초환	산수유환	어성초환	전화주시면 간략하게 설명드립니다
재료명 : 골담초분말, 쇠무릎뿌리, 황칠나무, 두충, 가시오피, 참당귀, 울금, 덩이뿌리, 쇠비름, 독활뿌리, 개다래열매, 천마뿌리, 율무, 질경이, 음나무줄기, 왕느릅나무껍질, 엉겅퀴, 산사나무열매, 맥아	재료명 : 산수유, 호박씨앗, 옥수수수염, 복분자, 사상자, 토사자, 구기자, 오미자, 천궁, 우슬, 복령, 산청목, 질경이, 뽕나무뿌리껍질, 두충, 산사열매, 맥아	재료명 : 어성초, 적하수오, 자소엽, 측백나무잎, 서목태, 원지, 우슬, 천문동, 마가목열매, 복분자, 솔잎, 뽕나무뿌리껍질, 녹차잎, 창출, 구기자, 산사나무열매, 맥아	
허리, 다리가 불편하십니까? 오래 걷기 힘드십니까? 계단 오르기 힘드십니까?	소변때문에 밤에 몇번씩 일어나십니까?	머리카락이 많이 빠지십니까?	

하지 · 아이케어센터 | 구입문의 : 02-969-9007 / HP.010-3245-0118
농협 : 301-0234-9633-61 / 예금주 : 윤제하

민속죽염 30년 전통
조상의얼과혼이담긴 우리민족의대표죽염

인체건강의 4대 요소는 “맑은 물, 맑은 공기, 맑은 음식, 질좋은 소금” 인데 이 중 “질 좋은 소금” 자색죽염은 총 아홉 번의 가열 과정을 거쳐 독성의 완전 제거와 약성의 완전 합성이 이루어지며 마지막 아홉 번째는 극강한 열에 의한 용융과정을 거쳐 탄생된 품질이 가장 우수한 죽염입니다. 이 죽염은 다음과 같은 기능이 있기에 일반인 및 환자들에게 꾸준한 섭취를 권장 할 수 있는 좋은 식품이라 할 수 있습니다.

① 강한 해독작용을 하며 간 기능을 좋게 한다.
② 소염작용을 하며 잇몸과 치아를 튼튼하게 해 준다.
③ 위장을 튼튼히 하며 식욕을 촉진시킨다.
④ 정혈작용 등 체질개선에 도움을 준다.



생명 에너지를 굳건히 지키면 약이 입에 이르기 전에 병이 낫는다.
환인동 요가마을

“죽염을 꾸준히 섭취하시면 인체의 자가 면역력이 높아져 감기 등 기타 질병 예방에 크게 도움이 됩니다”



◈ 9번 구운 민속죽염 ◈

가 루 자 죽 염	125g	45,000원
가 루 자 죽 염	230g	82,000원
알 갱 이 자 죽 염	125g	50,000원
알 갱 이 자 죽 염	230g	90,000원
휴대용 알갱이 자죽염	20g	9,000원

◈ 2번 구운 민속죽염 ◈
◆ 음식조리용 · 양치용 · 피부미용

보급형 민속죽염	500g	20,000원
	250g	10,000원

※ 단체 선물용으로 적합합니다. ※ 구입가 5만원 미만은 배송비 3천원 적용됩니다.

주문 및 상담 010-2823-0483
입금 : 농 협 150012-56-086122 예금주 : 김용희(민속죽염)
경북 영덕군 지품면 속곡길381-4 www.msjy.co.kr