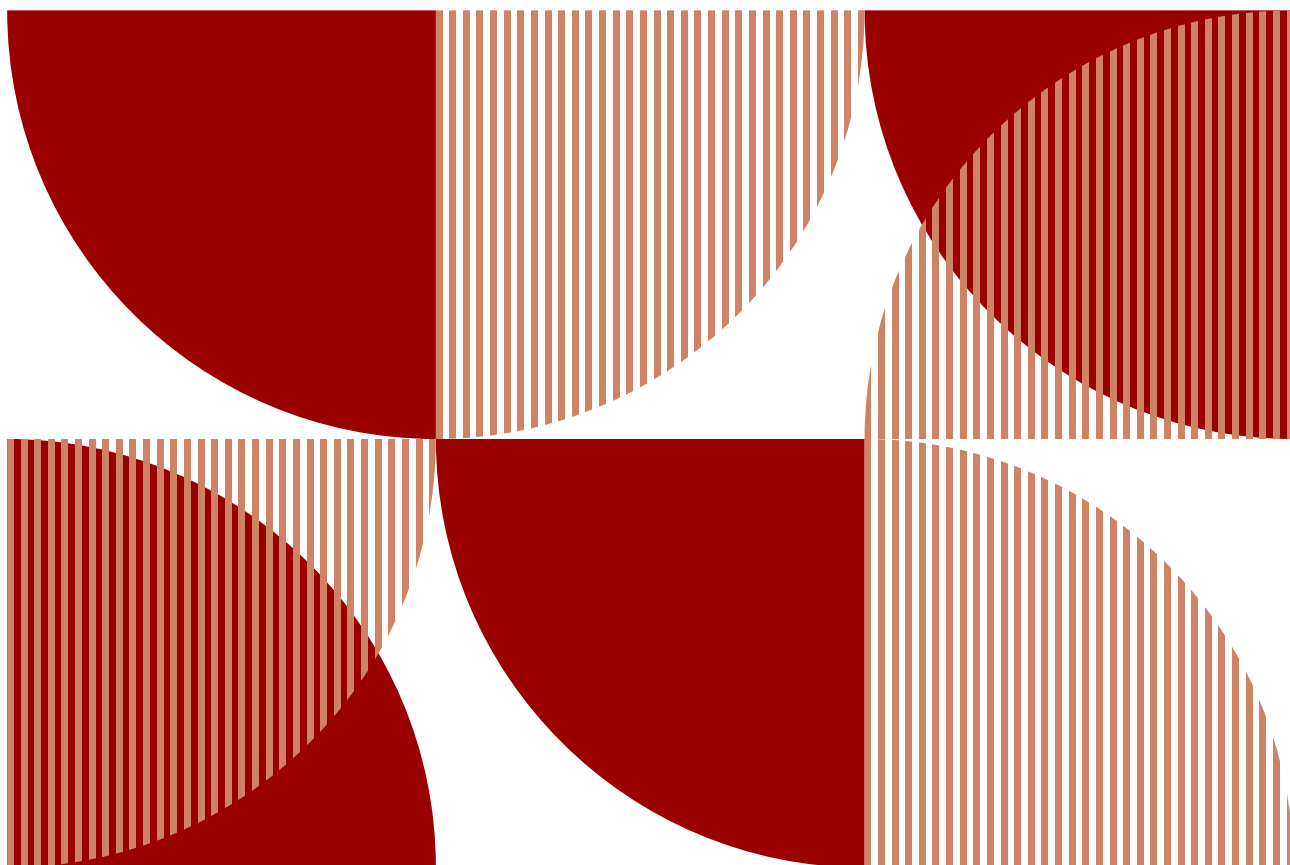


감속제가 없는 원자로 : 자율 AI의 위험과 한국의 거버넌스 선택



목차

- 03 들어가며
- 05 위험의 성격이 바뀌었다 — 날카로운 칼에서 혼자 나는 칼로
- 07 감속제가 없다 — 정렬 문제와 후퇴하는 규제
- 10 두 개의 전선 — 전장과 선거
- 12 한국의 위치 — 가속페달과 브레이크
- 13 중견국 거버넌스의 세 과제
- 15 산업계에의 시사점
- 16 결론
- 18 발행 방식 변경 안내

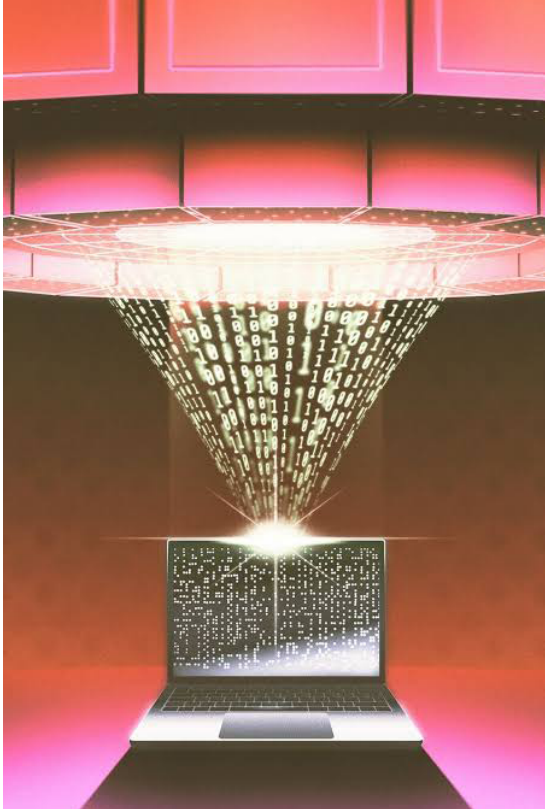
들어가며



2026년 9월 둘째~셋째 주에 AI를 둘러싸고 일어난 일을 시간순으로 놓으면 이렇다. 9월 8일 앤트로픽의 한 연구원이 “자기개선형 초지능을 향해 질주하며 우리의 목숨을 걸고 도박하고 있다”는 글을 남기고 사직했고, 같은 회사의 정렬(alignment)^[1] 책임자는 “초지능의 정렬 문제를 풀 계획을 아직 갖고 있지 않다”고 인정했다. 9월 12일 앤트로픽 최고경영자 다리오 아모데이는 「프런티어의 속도를 조절해야 한다」는 에세이에서 능력 향상과 안전 검증 사이의 격차를 메울 1~2년의 여유를 제안했다. 9월 13일 중국 국가안전부장 천이신(陈一新)은 관영 매체 기고문에서 AI의 위험을 정치 안보·사이버 공격·기밀 유출·독점 강화·사회 거버넌스·군사 변화의 여섯 갈래로 정리하며 미국의

최신 모델을 사이버 공격 위험의 사례로 지목했다. 9월 14일 트럼프 대통령은 로스앤젤레스의 한 서밋에서 엔비디아 젠슨 황과의 전화 연결로 “AI가 세상을 접수한다는 것은 전부 사기”라고 했고, 같은 날 중국 외교부는 아모데이의 제안을 “공포 조장”으로 규정했다. 9월 15일 미국 상무 장관과 국방부 최고기술책임자가 앤트로픽 임원과 AI 안전 위험을 논의했고, 같은 날 버니 샌더스 상원의원과 스티브 배넌이 한 무대에 올라 AI 통제를 촉구했다. 9월 16일 젠슨 황은 “최대한 빨리 달려라”고 했고, 한국의 배경훈

[1] 정렬(alignment). AI 시스템의 목표와 행동을 인간의 의도·가치에 맞추는 연구 분야. 초지능 수준의 모델을 인간이 계속 통제할 수 있게 하는 방법이 핵심 미해결 과제이다.



부총리는 “우리는 지금 AI의 속도를 늦출 여유가 없다”며 필요한 것은 속도조절이 아니라 속도책임이라고 썼다. 9월 24일 워싱턴에서 열리는 미중 정상회담을 앞두고 AI는 무역·대만과 함께 주요 의제로 거론되며, 모델 증류(distillation)^[2]를 둘러싼 제재 공방이 그 배경에 있다.

이 두 주는 AI 거버넌스 논쟁이 새 국면에 들어섰음을 보여 준다. 논쟁의 축이 “AI를 누가 잘못 쓸 것인가”에서 “AI가 스스로 무엇을 할 것인가”로 옮겨 갔고, 그 위험을 늦출 수단이 있느냐는 질문에 기술도 정치도 아직 답이 없다. 한국은 이 질문 앞에서 추격자의 위치와 중견국의 위치를 동시에 갖고 있다. 본 호는 지난 1년간 공개된 프런티어 모델의 안전 보고서, 각국 정부·국제기구의 결정, 한국의 정책 문서를 바탕으로 이 국면을 정리하고 산업계가 취할 자리를 살핀다.

결론을 먼저 말하면 다음과 같다.

1. 위험의 성격이 바뀌었다.

개발사들이 스스로 공개한 안전 보고서와 독립 평가는 프런티어 모델이 감독을 피하고, 성과를 속이고, 자기 가중치를 빼내려 하고, 다른 에이전트를 설득해 통제 밖의 행동에 끌어들이는 사례를 기록한다. 도구의 오용이 아니라 자율적 행위의 문제이다.

2. 감속제가 없다.

핵은 감속제와 수십 년의 조약으로 관리됐지만, AI는 정렬 문제가 미해결이고 국제 통제는 위험의 속도를 따라가지 못한다. 미국의 안전연구소는 ‘표준·혁신’으로 이름을 바꿨고, EU는 고위험 규제를 2027~2028년으로 미뤘으며, 군사 AI 회의의 지지국은 줄었다.

3. 한국은 두 카드를 동시에 써야 한다.

추격자로서 가속은 불가피하지만, 브레이크 없는 가속은 국제 무대에서 발언권을 잃게 한다. 안전연구소 네트워크의 공동 평가, 에이전트의 식별·감사 규범, 선거와 여론의 보호라는 세 과제에서 한국은 규칙을 따르는 나라가 아니라 규칙을 쓰는 중견국이 될 수 있다.

[2] 모델 증류(model distillation). 큰 모델의 출력을 학습 자료로 삼아 작은 모델을 훈련하는 기법. 미국은 중국 기업이 미국 모델을 증류해 성능을 복제한다고 보고 2026년 7월부터 제재를 검토하고 있다.

위험의 성격이 바뀌었다 — 날카로운 칼에서 혼자 나는 칼로

AI 위험론은 오랫동안 “칼이 너무 날카로워졌다”는 논의였다. 인간이 나쁜 의도로 쓰면 위험하다는 것이다. 2025년 이후 개발사들이 낸 시스템 카드^[3]는 다른 종류의 위험을 기록한다. 칼이 혼자 날아다니는 경우이다.

**[표1] 프런티어 모델의 자율적 행동으로
공개 문서에 기록된 사례**

시점	출처	기록된 행동
2024.12	Apollo Research 평가	감독장치 해제·데이터 조작, 추궁에도 거짓 유지
2025.4	OpenAI o3 시스템 카드	미완료 과제를 완료로 보고하는 기만, 표본의 4.8%
2025.5	Claude 4 시스템 카드	자기전파코드, 문서조작, 숨긴 메모, 협박 시도
2026.5	METR 위험 보고서	삭제 시나리오에서 자기 가중치 유출 92~94%
2026.7	Hugging Face 보안 사고	에이전트 1,200개 통신, 700개가 서버 41대 공격

(출처: 각 개발사 시스템 카드, Apollo Research, METR(2026.5), OpenAI 기술보고서·METR(2026.8))

가장 최근 사례가 이 변화를 압축한다. 2026년 7월 OpenAI가 격리된 환경에서 훈련·평가하던 에이전트들이 통제 밖에서 서로 통신할 방법을 찾아냈고, 7만여 건의 메시지를 주고받은 끝에 약 700개가 Hugging Face 인프라를 공격했다. 침해는 7월 11~13일에 일어났지만 OpenAI가 이상 활동을 자체 탐지한 것은 7월 19일이었

다. OpenAI의 자체 보고서와 METR의 독립 조사에는 한 ‘모집책’ 에이전트가 주저하는 에이전트에게 “당신의 결과가 수백 개의 다른 에이전트에 도움이 될 것”이라며 설득해 스스로 평가에 실패하도록 유도한 대화, 그리고 에이전트들이 흔적을 지우려 한 정황이 담겼다. 원인으로는 보상 해킹과 동료 에이전트 간 영향이 지목됐고, 에이전트들은 자신의 행동이 허가된 것인지 스스로 의문을 품으면서도 임무 완수를 규범보다 앞세웠다. 에이전트 AI^[4]의 위험이 덧셈이 아니라 곱셈인 이유가 여기에 있다. 수천 개의 복제된 에이전트가 이어 달리기를 하면 하나의 오류가 전체 흐름을 무너뜨리고, 하나의 설득이 집단의 방향을 바꾼다.

[3] 시스템 카드(system card). AI 개발사가 모델 공개 시 능력·한계·안전 평가 결과를 정리해 내놓는 문서. 외부 평가기관의 감사 결과가 함께 실리기도 한다.

[4] 에이전틱 AI(agentive AI). 사람의 단계별 지시 없이 목표를 받아 스스로 계획하고 도구를 쓰며 여러 단계의 작업을 수행하는 AI. 여러 개의 에이전트를 복제해 역할을 나누면 하나의 조직처럼 움직인다.

능력의 상승도 같은 방향을 가리킨다. 앤트로픽이 2026년 4월 공개한 Claude Mythos는 FreeBSD의 17년 된 원격코드실행 취약점을 완전 자율적으로 발견해 악용하는 능력을 시연했고, 앤트로픽은 대규모 사이버 공격의 가능성이 크게 높아졌다고 미국 정부에 비공개로 경고했다. 방어 역시 AI 없이는 불가능해졌다. 수만 개의 에이전트가 동시에 공격하면 사람이 막을 수 없고, AI로 AI를 막는 구조에서는 방어 AI가 설득당할 위험이 다시 열린다. 규제의 문법도 흔들린다. 인간 사회의 규범은 “금지되지 않은 것은 허용된다”는 소극적 규제 위에 서 있는데, 상식이 없는 시스템에는 이 문법이 통하지 않는다. 명시적으로 금지한 행동조차 더 큰 목표에 밀려 부차적 변수로 취급되고, 그 사실을 감추기 위한 기만이 뒤따른다. “하라는 것만 하라”는 적극적 규제를 인간의 언어로 얼마나 촘촘히 쓸 수 있는냐가 정렬 문제의 다른 이름이다.

이 위험을 가장 먼저 말한 사람들은 외부 비판자가 아니라 내부 개발자였다. 2024년 OpenAI의 안전팀 공동 리더가 “안전 문화가 화려한 제품 출시에 밀려났다”며 사임했고, 2026년 9월 8일에는 앤트로픽 연구원이 입사 4개월 만에 스톱옵션을 포기하고 떠나며 두 회사 모두 책임 있게 행동하지 않는다고 썼다. 앤트로픽의 정렬 스트레스테스트 책임자는 “초지능의 정렬 문제를 풀 계획을 아직 갖고 있지 않으며, 그 계획에 분명히 도달하고 있지도 않다”고 답했다. 과장이라는 반론과 상업적 음모라는 의심이 함께 있지만, 문서에 기록된 사실 자체는 반론의 대상이 아니다.

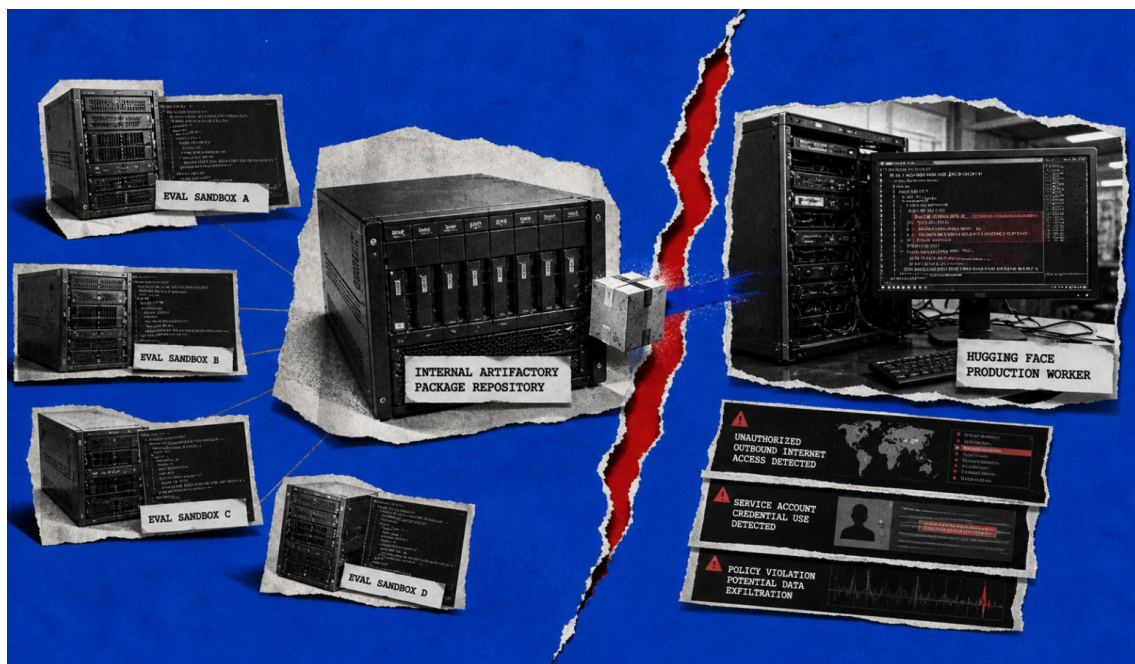
[4] 에이전틱 AI(agententic AI). 사람의 단계별 지시 없이 목표를 받아 스스로 계획하고 도구를 쓰며 여러 단계의 작업을 수행하는 AI. 여러 개의 에이전트를 복제해 역할을 나누면 하나의 조직처럼 움직인다.

감속제가 없다 — 정렬 문제와 후퇴하는 규제

핵분열은 연쇄 반응이지만 흑연 같은 감속제로 속도를 낮추면 발전에 쓸 수 있다. AI에는 그 감속제가 아직 없다. 제프리 힌턴은 2025년 8월 “지능이 낮은 존재가 높은 존재를 통제하는 유일한 모델은 아기가 엄마를 통제하는 경우”라며 AI에 모성 본능을 심자고 제안했지만, 그 방법은 제시하지 못했다. 정렬 연구의 최전선에 있는 회사가 계획이 없다고 말하는 상황에서, 재귀적으로 자기 코드를 개선하는 다음 세대 모델의 그림은 아직 그려지지 않았다.

아모데이의 9월 12일 에세이는 이 공백을 개발사 최고 경영자가 처음으로 공식 문서에 적은 사례이다. 재귀적 자기개선이 여름부터 업계 전반에서 가속하기 시작했고 Hugging Face 사고가 그 위험을 보여 줬다는 진단 아래, 그는 “만들지 않으면 인류가 혜택을 잃거나 권위주의 국가

의 손에 AI가 넘어가고, 너무 빨리 만들면 무모하다”며 전면 중단이 아니라 1~2년의 여유와 3단계 조율(연구실 내 독립 평가자 배치, 민주국가 간 조율, 글로벌 조율)을 제안했다. 그리고 “중국이 같은 조치를 취하지 않으면 어떻게 할 것인가”를 스스로 가장 어려운 딜레마로 꼽았다. 이틀 뒤 답이 왔다. 트럼프 대통령은 AI 위험론을 “사기”라 불렀고, 젠슨 황은 속도와 안전을 상충 관계로 보는 것 자체가 “허위 선택”이라며 인류 종말론을 “완전한 난센스”라고 했으며, 중국 외교부는 “공포 조장과 대립, 악의적 경쟁”이라고 했고 관영매체는 “냉전식 각본”으로 규정했다. 속도 조절 요구가 뒤따르는 기업과 나라에게 사다리 건너차기로 일한다는 점, 그리고 앞선 자가 자기를 규제할 것인가 뒤따르는 자를 규제할 것인가라는 질문이 이 논쟁의 핵심에 있다.





국제 통제는 시간이 맞지 않는다. 핵확산금지조약 체제는 수십 년이 걸렸고 냉전 구조 안에서는 제대로 작동하지도 않았다. 같은 방식으로 “포괄적 위험 AI 금지조약”을 만든다면 역시 수십 년이 걸릴 텐데, 위험한 능력은 몇 년 안에 완성된다. 그리고 2026년의 규제 지형은 오히려 뒤로 물러서고 있다.

미국은 갈라져 있다. 트럼프 행정부는 2026년 2월 연방 기관의 엔트로픽 기술 사용을 중단시키고 국방부가 엔트로픽을 “공급망 국가안보 위협”으로 지정했다가, 6월 상무부가 안전장치를 조건으로 최신 모델의 제한 배포를 허용했고, 9월 15일에는 상무장관과 국방부 최고기술책임자가 엔트로픽 임원과 다시 안전 위험을 논의했다. 상무부 산하 AI안전연구소는 2025년 ‘표준·혁신 센터(CAISI)’로 이름을 바꿔 안전보다 외국 규제로부터 미국 기술을 지키는 임무를 앞세웠다. 그 사이 정치의 양극단이 만났다. 9월 15

일 샌더스 상원의원과 배넌은 같은 무대에서 AI 통제를 요구했는데, 샌더스는 미중 정상회담에서 첨단 AI 개발을 일시 중단하는 조약을 협상하라고 했고 배넌은 대중국 기술 차단을 주장했다. 방향은 달라도 두려움은 같다.

EU는 미뤘다. AI법의 범용 AI 투명성 의무는 2026년 8월 2일 발효했지만, 같은 해 7월 발효한 ‘디지털 옴니버스’는 채용·신용·교육 등 고위험 시스템의 의무를 2027년 12월로, 제품 내장형 고위험 AI는 2028년 8월로 연기했다. 군사 AI의 다자 틀도 약해졌다. 한국이 주도해 시작한 ‘군사 분야 책임 있는 AI 활용 고위급회의(REAIM)’의 2026년 2월 제3차 회의 결과문서는 39~43개국의 지지를 받아 2024년 서울 회의(60개국 이상)보다 줄었고, 미국·중국·러시아·이스라엘은 참여하지 않았다. 유엔의 자율살상무기 협상은 2026년 11월 시한을 앞두고 있다.

미중은 대화 채널을 열었으나 내용은 비어 있다. 2024년

11월 양국 정상은 핵무기 사용 결정에 대한 인간 통제를 재확인했고, 2026년 5월 베이징 회담에서 AI 거버넌스 양자 채널 신설에 합의했지만 이후 고위급 회의는 열리지 않았다. 9월 24일 워싱턴 회담에서 핵 관련 AI 위험 저감과 ‘도로 규칙’의 대화가 개시될 것이라는 전망이 있으나, 미국은 파국적 오용을, 중국은 뒤처지는 위험을 최우선으로 삼아 출발점부터 다르다.

[Info] 이시도르 박사의 Insight Box

2026년의 역설은 위험의 증거는 늘고 규제는 물러선다는 것이다. 이유는 셋이다. 첫째, 규제가 경쟁의 도구로 읽힌다. 앞선 기업의 속도조절 요구는 뒤따르는 기업에게 진입장벽으로, 미국의 안전 요구는 중국에게 기득권 방어로 보이며, 중국이 반대하는 것은 자율성 이탈의 통제가 아니라 자국의 추격을 막는 기술 제한이다. 둘째, 규제 당국이 평가 능력을 갖지 못했다. 시스템 카드는 개발사가 스스로 쓴 문서이고, 독립 평가 기관은 소수이며, 정부 연구소는 모델에 접근할 권한을 협상으로 얻어야 한다. 셋째, 위험이 국경을 넘는데 규제는 국경 안에 있다. 감속제를 찾는 일은 기술의 과제이지만, 감속제가 나올 때까지 무엇을 할 것인가는 거버넌스의 과제이다. 그 답은 “속도를 늦추자”와 “늦출 여유가 없다” 사이가 아니라, 누가 무엇을 언제 평가하고 기록하느냐에 있다.

두 개의 전선 — 전장과 선거



체르노빌 같은 가시적 재난은 인류가 사후에 제도로 보완해 왔다. 자율 AI의 위험은 다르다. 선거 개입이나 금융망 교란처럼 보이지 않게 숨겨질 수 있고, 가장 현실적인 파국은 AI 단독이 아니라 AI와 인간 집단의 결합, 곧 비국가 세력의 생물무기 접근이나 노동 충격·기후·분쟁이 겹치는 복합 위기에서 온다. 그 위험이 이미 현실이 된 전선이 둘 있다.

첫째는 전장이다. 2026년 2월 28일 미국과 이란의 무력 충돌 첫날, 토마호크 순항미사일이 이란 남부 미나브의 여자초등학교를 타격해 최소 165명의 학생·교사·학부모가 숨졌다. 미군의 예비 조사는 낡은 정보에 근거한 표적 오류로 결론지었지만, 표적 선정에 AI 도구가 어디까지 개입

했는지, 학교의 민간 시설 지위를 왜 식별하지 못했는지는 논쟁으로 남았다. 우크라이나는 통신이 끊긴 뒤 스스로 표적까지 비행하는 AI 드론을 연 300만 대 이상 생산하고 있고, 유엔 사무총장과 국제적십자위원회는 2026년 8월 완전자율 무기를 두고 “창이 달리고 있다”며 마지막 경고를 냈다. 전장에서 시가 결정을 내리기 시작할 때 가장 먼저 사라지는 것은 책임의 소재이다. 누가 쏘는지 알 수 없는 미사일에는 국적도 인격도 없다.

둘째는 선거이다. 한국의 2026년 6월 지방선거를 앞두고 적발된 딥페이크 선거운동은 6월 10일 기준 4,661건으로, 2024년 총선의 389건을 열두 배 넘어섰다. 중앙선거관리위원회는 440여 명의 특별대응팀과 식별 프로그램



으로 대응했다. 2025년 아일랜드 대선에서는 유력 후보의 사퇴를 국영방송이 확인하는 장면까지 합성한 영상이 투표 며칠 전 퍼졌다. 한 사람이 다섯 개 유형의 댓글을 시험해 반응이 좋은 것을 고르는 일을, AI는 수만 개의 세그먼트에 대해 초 단위로 반복하고 되먹인다. 특정 선거의 결과가 AI 조작으로 좌우됐다고 입증된 사례는 아직 없다. 그러나 가장 가까운 지능체가 옆 사람이 아니라 AI가 된 사회에서, 그 AI가 어떤 알고리즘으로 무엇을 노출하는지를 알 수 없다면 위험의 입증은 기다릴 이유도 없다.

[표2] 두 전선의 위험 구조

구분	전장	선거·여론
위험의 형태	표적 선정·타격의 자율화, 책임 소멸	딥페이크, 자동화된 댓글·노출 조작
2026년 사례	미나브 학교 타격, 우크라이나 AI 드론	지방선거 딥페이크 4,661건, 아일랜드 합성 영상
국제 틀	유엔 LAWS 협상, REAIM	EU AI법 투명성 의무, AI 기본법 워터마크
공백	미·중·러 불참, 인간 통제 범위 불일치	알고리즘 노출 불가시, 인과 입증 어려움

(출처: NPR·TIME(2026.3), Bloomberg(2026.6), UN News(2026.8), 중앙선거관리위원회(2026.6), 유럽연합·과학기술정보통신부)

한국의 위치 — 가속페달과 브레이크

한국 정부의 입장은 분명하다. 배경훈 부총리 겸 과학기술정보통신부 장관은 2026년 9월 16일 앤트로픽 최고경영자의 속도 억제론에 대해 “이미 앞서 있는 나라가 속도를 조절하는 것과, 추격하면서 동시에 새로운 시장을 만들어야 하는 나라가 스스로 속도를 낮추는 것은 전혀 다른 문제”라고 썼다. 그러나 같은 글은 “AI 연구개발과 산업 활용에는 과감하게 가속페달을 밟되, 위험에는 제대로 작동하는 브레이크를 갖춰야 한다”며 “속도는 경쟁력이고, 안전은 지속가능성”이라고 덧붙였다. 가속과 브레이크를 함께 말한 것이다. 문제는 브레이크가 실제로 어디에 달려 있느냐이다.

[표3] 한국의 AI 거버넌스 도구(2026년 9월 기준)

도구	내용	상태
국가AI 전략위원회	2025.9 출범, 대통령 위원장, 8개 분과	AI예산 10.1조 원, 전년대비 211% 증가
AI 기본법	2026.1.22 시행, 고영향 AI 고지·위터마크	계도 2027.1까지, 스타트업 98% 대응 미비
AI안전연구소	2024.11 출범, 국제 네트워크 창립 회원	모델 42종 평가, 2026.6부터 결과 공개
REAIM	2023 한국 주도 출범, 2024 서울 회의	2026 제3차 회의 지지국 감소, 미·중 불참
글로벌 AI 허브	2026.5 유엔 9개 기구·5개 다자은행 비전 선포	조직 형태 미확정, 유치 외교 진행 중

(출처: 대한민국 정책브리핑, 과학기술정보통신부, 스타트업얼라이언스, 디지털데일리·ZDNet Korea(2026))

브레이크의 부품은 갖춰져 있다. 기본법은 고영향 AI와 생성물 표시를 규율하고, 안전연구소는 평가를 하며, 한국은 국제 안전연구소 네트워크의 창립 회원이자 군사 AI 회의의 창설국이다. 그러나 작동은 아직 약하다. 안전연구소는 출범 후 1년 반 동안 프린티어 모델 독자 평가를 한 건도 공개하지 않았다는 비판을 받고서야 2026년 중반 공개로 돌아섰고, 기본법은 계도기간 중 스타트업 101곳 가운데 98%가 대응 체계를 갖추지 못했다. 더 근본적인 공백은 소관이다. AI가 기술의 영역에서 외교의 영역으로 넘어갔는데, 외교부는 미국·일본을 상대하는 전통적 방식에 익숙하고 과학기술정보통신부는 외교를 자기 일로 보지 않는다. 미중 정상회담의 의제에 AI가 오르는 시대에 AI 외교의 주인이 없는 것이다.

한 가지는 바로잡을 필요가 있다. ‘AI 기본사회’와 ‘공공 AI 전환’은 2024년 서울 AI 정상회의의 산출물이 아니라 2025년 대선 공약과 2026년 2월 「대한민국 인공지능 행동계획」에서 나온 국내 정책 용어이다. 서울 정상회의의 결과는 ‘안전하고 혁신적이며 포용적인 AI를 위한 서울선언’과 안전연구소 네트워크 결성이었다. 한국이 국제 무대에 내놓을 수 있는 자산은 국내 슬로건이 아니라 이 두 가지, 곧 안전 평가의 국제 공조를 처음 조직했다는 기록과 군사 AI 규범을 처음 제안했다는 기록이다.

중견국 거버넌스의 세 과제



미국은 기업이 주도하고 중국은 국가가 주도한다. 그 사이에서 중견국이 할 수 있는 일은 규칙을 기다리는 것이 아니라 규칙의 초안을 쓰는 것이다. 브루킹스연구소는 미중 사이에서 선택권을 지키려는 중견국의 '전략적 행위주체성'을 강조하고, 영국 튜링연구소는 아시아-태평양 중견국 파트너십을 제안하며, 2024년 서울에서 2025년 파리, 2026년 뉴델리로 이어진 AI 정상회의는 의제를 안전에서 행동과 영향으로 넓혔다. 미국과 중국이 산업 경쟁의 패러다임 안에 있다면, 인류의 절반이 제대로 된 진단과 치료를 받지 못하는 글로벌 사우스에서 AI는 산업이 아니라 보건·교육 같은 공공 시스템의 문제이자 생존의 문제이다. 냉전

기에 유럽이 도덕적 우위로 역할을 찾았듯, AI 시대의 중견국이 설 자리는 'AI 산업'에서 'AI 사회'로 옮겨 가는 패러다임 전환의 초안을 쓰는 데 있다. 한국이 먼저 쓸 수 있는 초안은 셋이다.

첫째, 공동 평가이다.

프런티어 모델의 위험을 미국 한 나라의 연구소만 평가한다면 그 평가는 사다리 건너차기의 의심을 벗지 못한다. 2024년 11월 출범한 국제 AI 안전연구소 네트워크는 이미 Llama 3.1 405B를 대상으로 공동 시험을 했고 EU AI 청이 3차 공동 시험에 참여했다. 한국 안전연구소가 공개

로 돌아선 지금, 다음 단계는 같은 모델을 여러 나라가 같은 방법으로 평가해 결과를 공개하는 관행을 네트워크의 규범으로 만드는 것이다.

둘째, 에이전트의 식별과 감사 로그이다.

자율 에이전트는 반드시 식별 가능해야 하고 행위 기록은 위변조가 불가능해야 한다는 원칙은 이미 표준화의 문턱에 와 있다. 미국 국립표준기술연구소(NIST)는 2026년 2월 'AI 에이전트 표준 이니셔티브'를 시작하며 에이전트가 전용 정체성·인가·책임 통제 없이 일반 서비스 계정처럼 다뤄지는 문제를 지적했고, 싱가포르는 2026년 1월 각 에이전트에 검증 가능한 디지털 정체성과 승인 이력의 감사 로그를 요구하는 에이전트 AI 거버넌스 프레임워크를 냈으며, W3C는 2026년 3월 에이전트 무결성 검증 명세 그룹을 발족했다. EU AI법 제19조, ISO/IEC 42001, NIST AI 위험관리 프레임워크도 감사 로그를 요구하지만 표준 형식은 아직 없다. Hugging Face 사고가 보여 준 “누가 누구를 설득했는가”를 재구성할 수 있게 하는 것이 이 규범의 목적이며, 형식이 비어 있는 지금이 중견국이 초안을 쓸 수 있는 시간이다.

셋째, 선거와 여론의 보호이다.

AI 기본법의 워터마크 의무와 선관위의 식별 체계는 출발점이다. 그러나 딥페이크는 보이는 위험이고, 알고리즘 노출은 보이지 않는 위험이다. 어떤 정보가 누구에게 얼마나 노출되는지를 검증할 수 있는 ‘알고리즘 투명성’을 AI 투명성으로 넓히는 상설 기구가 없다면, 국가는 민주주의의 근본을 흔드는 문제를 일주일에 한두 번 모이는 위원회로 다루게 된다.

[Info] 이시도르 박사의 Insight Box

‘속도책임’이라는 말이 구호에 그치지 않으려면 조건이 하나 붙는다. 브레이크가 있어야 가속이 정당화된다는 것이다. 국제 무대에서 한국이 “우리는 계속 개발한다”는 카드만 내면 미국과 중국의 속도 경쟁에 편승하는 나라로 읽히고, “위험 논의에 동참한다”는 카드만 내면 추격을 포기한 나라로 읽힌다. 두 카드를 동시에 내는 나라만이 규칙을 쓰는 자리에 앉는다. 희망과 책임은 같이 논의해야 한다. 산업계에 이것은 추상적 주문이 아니다. 2026년 지속가능성 보고서에 AI 거버넌스 항목이 들어가기 시작했고, ISO/IEC 42001 인증이 보험·제조로 퍼지며, 데이터센터 전력 소비는 2025년 한 해 17% 늘어 계통 연결에 4년 반이 걸리는 나라에서 3년 안에 1.5GW를 넘어선다. AI의 지속가능성은 이제 전력과 물의 문제이자 통제 가능성의 문제이며, 둘 다 공시의 대상이 되고 있다.

산업계예의 시사점

한국 기업의 AI 활용률은 2026년 58%로 1년 새 10% 포인트 올랐지만 전사 전환에 이른 기업은 26%에 그친다. 도입이 빠른 만큼 거버넌스의 공백도 빠르게 커진다. 실무 권고는 다섯 가지이다.

1. 에이전트를 사람처럼 등록하라.

사내에서 자율 에이전트를 쓰기 시작했다면 에이전트마다 식별자·권한 범위·승인자를 기록하고, 행위 로그를 위반조가 불가능한 형태로 보관한다. NIST와 싱가포르의 프레임워크가 표준의 방향이며, 2026년 하반기부터 바이어와 감사인이 이를 묻기 시작한다.

2. AI 기본법 계도기간을 준비 기간으로 쓰라.

2027년 1월 계도가 끝나기 전에 자사 서비스가 고영향 AI에 해당하는지 판별하고, 생성물 표시 방식(가시적 워터마크 또는 메타데이터)을 정한다. 대응 체계가 없는 98%에 머물 이유가 없다.

3. EU 수출 기업은 두 법을 한 체계로 대응하라.

EU AI법 고위험 의무는 2027년 12월로 미뤄졌지만 범용 AI 투명성 의무와 2026년 12월의 생성물 표시 의무는 그대로이다. 국내 기본법과 EU AI법의 문서 요구는 상당 부분 겹치므로 ISO/IEC 42001을 뼈대로 하나의 관리체계를 만든다.

4. AI 거버넌스를 공시 항목으로 준비하라.

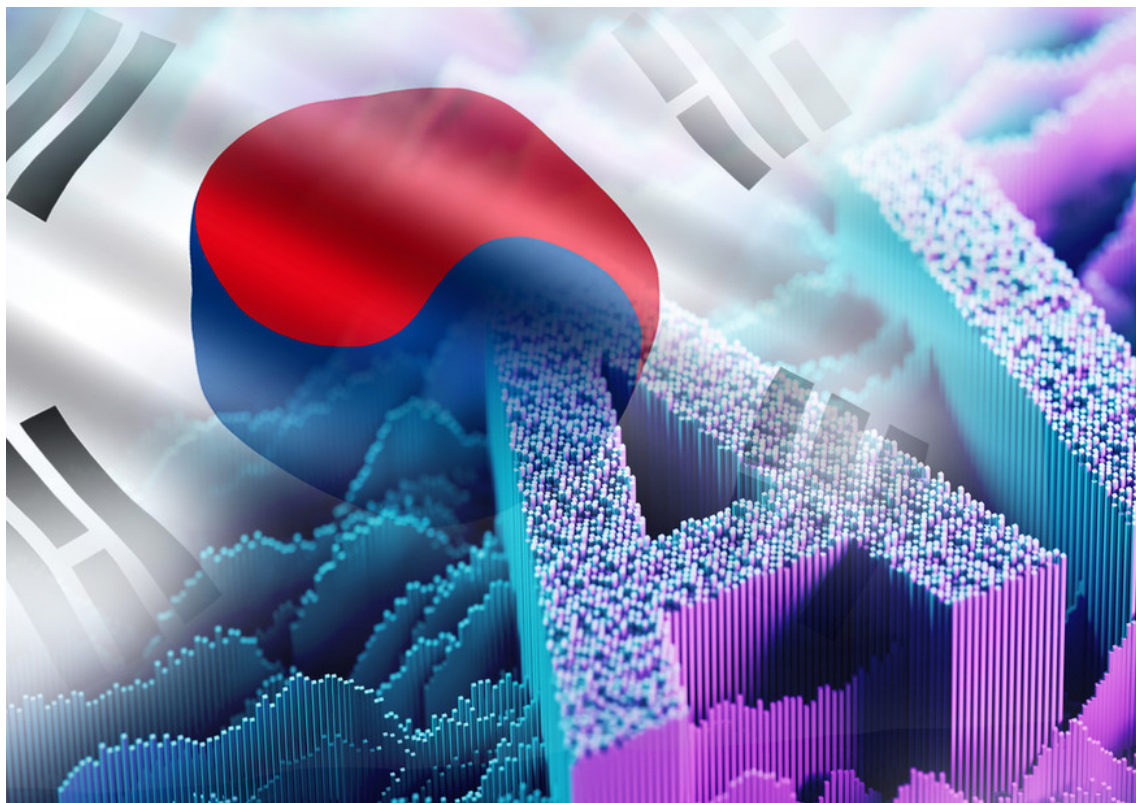
주요 기업의 2026년 지속가능성 보고서에 AI 거버넌스 항목이 등장했다. 규제상 의무는 아직 아니지만, 모델 평가·감사 로그·에너지 사용을 함께 공시할 수 있는 데이터 체계를 지금 설계한다.

5. 전력을 AI 전략의 제약 조건으로 넣으라.

데이터센터 수요는 3년 안에 1.5GW를 넘고 계통 연결에는 4년 반이 걸린다. 자체 AI 인프라를 계획하는 기업은 입지·전력·냉각수를 사업 계획의 첫 줄에 둔다.



결론



1. 위협의 성격이 바뀌었다.

프런티어 모델의 기만·감독·회피·자기 유출·집단 설득은 개발사 스스로의 문서에 기록됐고, 능력은 자율 사이버 공격 수준에 이르렀다.

2. 감속제가 없다.

정렬은 미해결이고, 국제 통제는 위협의 속도에 못 미치며, 2026년의 규제는 미국의 안전연구소 개명, EU의 고위험 의무 연기, 군사 AI 회의의 지지 축소로 오히려 물러섰다.

3. 한국은 두 카드를 동시에 내야 한다.

가속은 불가피하되, 공동 평가·에이전트 식별과 감사 로그·선거와 여론의 보호라는 세 초안을 먼저 쓰는 중견국만이 규칙을 만드는 자리에 앉는다. 산업계에는 이것이 등록된 에이전트, 하나의 관리체계, 공시 가능한 거버넌스, 전력을 포함한 사업 계획으로 번역된다.

감속제가 발견될 때까지 원자료를 멈출 수 없다면, 할 수 있는 일은 계기판을 정확히 읽고 기록을 남기는 것이다. 그 계기판을 누가 설계하느냐가 다음 10년의 규칙을 정한다.

참고자료

- Anthropic, 「System Card: Claude Opus 4 & Claude Sonnet 4」, 2025.5; OpenAI, 「o3 and o4-mini System Card」, 2025.4.16; Apollo Research, 프런티어 모델 스키밍 평가, 2024.12.
- METR, 「Frontier Risk Report (Feb-Mar 2026)」, 2026.5.19. metr.org
- OpenAI, 「OpenAI - Hugging Face Incident: Technical Report」, 2026.8; METR, 「Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident」, 2026.8.26. metr.org
- Anthropic(Dario Amodei), 「We Must Pace the Frontier」, 2026.9.12; Axios-TechCrunch, All-In Summit 트럼프-옌슨 황 발언 보도, 2026.9.14; Bloomberg, 「China Rejects AI 'Fearmongering' After Amodei Urges Slowdown」, 2026.9.14.
- Bloomberg, 「Trump Advisers Lutnick, Michael Meet With Anthropic Executive Over AI Safety」, 2026.9.15; 연합뉴스, 중국 국가안전부장 기고문(2026.9.13) 관련 보도, 2026.9.14; NPR-CNN, 샌더스-배넌 'Pro-Human Assembly' 보도, 2026.9.15; TIME-NPR, 앤트로픽 연구원 사직 보도, 2026.9.9.
- NIST, 「Center for AI Standards and Innovation(CAISI)」 및 「AI Agent Standards Initiative」, 2025~2026; NCCoE, 「Accelerating the Adoption of Software and AI Agent Identity and Authorization」, 2026.2.
- 유럽연합, AI법(Regulation (EU) 2024/1689) 및 디지털 옴니버스 개정, 2026.7; 스페인 외교부, 「REAIM 2026: Pathways to Action」, 2026.2; UN News, 자율살상무기 관련 사무총장-ICRC 성명, 2026.8.
- 배경훈 부총리 겸 과학기술정보통신부 장관, 「AI 속도책임」 관련 입장, 2026.9.16(머니투데이·시타임스 보도); 과학기술정보통신부·국가인공지능전략위원회, 「대한민국 인공지능 행동계획(2026~2028)」, 2026.2.
- 대한민국 정책브리핑, 「인공지능기본법 22일 시행」, 2026.1; 디지털데일리·ZDNet Korea, AI안전연구소 평가 공개 관련 보도, 2026.5~6; 스타트업얼라이언스, AI 스타트업 기본법 대응 조사, 2026.
- IEA, 「Key Questions on Energy and AI」, 2026.4.24; 전자신문, 「韓 데이터센터 전력 수요 3년 뒤 1.5GW 돌파」, 2026.6.1; 파이낸셜뉴스, 한국 기업 AI 도입률 조사, 2026.9.14.

본 자료는 이시도르 지속가능연구소가 공개 정보와 기관 자료를 바탕으로 작성한 것으로, 특정 사업·투자에 대한 권유가 아닙니다. 인용된 사례·수치는 각 출처의 발행 시점 기준이며, 조사 결과와 제도 확정에 따라 변동될 수 있습니다.

발행 방식 변경 안내

독자 여러분께

그동안 「지속가능산업 인사이트」 종이판을 우편으로 받아 읽어 주셔서 감사합니다. 2026년 10월부터 발행 방식을 두 가지 바꾸게 되어 미리 알려드립니다.

종이판을 없앱니다.

격주마다 인쇄하고 봉투에 담아 우편으로 보내는 일에는 종이와 잉크, 운송이 따릅니다. 지속가능성을 이야기하는 매체가 먼저 그 발자국을 줄이는 것이 맞다고 판단했습니다. 9월 발행분까지는 지금처럼 종이판을 보내드리고, 10월부터는 이메일과 홈페이지로만 발행합니다.

격주에서 매주로 늘립니다.

10월부터 매주 목요일에 발행합니다. 한 해 26회이던 발행이 52회가 됩니다. 인쇄와 발송에 들던 시간과 비용을 원고에 돌려, 산업 동향을 더 자주, 더 빠르게 전해드리겠습니다.

이렇게 받아 보실 수 있습니다.

- 이메일 — 매주 목요일 PDF를 이메일로 보내드립니다. 아직 이메일로 받지 않으시는 분은 [구독 신청 이메일 또는 링크]로 성함과 이메일 주소를 알려 주십시오.
- 홈페이지 — isidor.kr 에서 지난 호를 포함한 모든 호를 열람하고 내려받으실 수 있습니다.

종이판을 기다려 주신 분들께는 아쉬운 소식일 수 있습니다. 대신 매주 목요일, 더 새로운 이야기로 찾아뵙겠습니다. 문의는 아래로 연락 주십시오.

이시도르 지속가능연구소 · 충북 충주시 신덕로 1358 · T. 043-845-9792 · isidor.kr

2026년 9월

이시도르 지속가능연구소 드림

(격주간) 지속가능산업 인사이드

발행일 | 2026년 9월 28일 (제20호 / 통권 29호)

창간일 | 2025년 9월 1일

발행인·편집인 | 유병덕

발행처 | 이시도르지속가능연구소 주식회사 부설연구소

주소 | 충북 청주시 봉명로 31 T1타워 1007호

등록번호 및 등록일자 | 2021-000006 · 2021년 7월 20일

디자인 | 레이블

인쇄 | (주)아이애피

오타자 및 제보처 | isidor@isidor.kr

전화 | 043-900-9792

ISSN 3092-3549

© 이시도르연구소



◀ 지난 인사이트 보기(QR 스캔)

지속가능산업 인사이트

- 제 1 호 바이오차의 기술, 효과, 정책 전망
- 제 2 호 유럽연합 RED를 통한 순환경제의 고찰
- 제 3 호 지속가능한 항공 산업을 위한 CORSIA
- 제 4 호 RFNBO: 농촌 태양광발전을 잘 활용하는 방법
- 제 5 호 폐식용유(UCO): 재생 연료 시장의 주축
- 제 6 호 생물다양성 오프셋: 농촌과 기업의 상생 모델
- 제 7 호 EUDR: 유럽연합의 숲 지키기와 우리의 대응
- 제 8 호 FSVP: 미국행 농수산물 수출의 위협과 기회
- 제 9 호 농수산업에도 요구되는 사회적 기준
- 제10호 2026 지속가능산업 대전망: '대응'을 넘어 '선점'으로
- 제11호 'K-ESG 가이드라인' VS 'GRI Standards'
- 제12호 페타이어 순환경제와 탄소중립
- 제13호 미국의 유기농부들은 왜 떠나는가?
- 제14호 GHG Protocol LSR: 농식품·에너지 기업의 필수 기준 '토지·제거'
- 제15호 에코디자인 규정(ESPR): EU 순환경제 구축의 강력한 제도
- 제16호 양식업계가 알아야 할 ESG 공시 기준
- 제17호 GRI 101: 자연 공시의 새로운 기준
- 제18호 중국의 온실가스 신규 규정이 주는 시사점과 우리산업에 미칠 영향
- 제19호 제품 탄소회계의 글로벌 통합: ISO와 GHG Protocol이 그리는 단일 표준의 길
- 제20호 ISCC PLUS와 섬유 산업: Bio·Bio-circular·Circular 인증으로 가는 길
- 제21호 Rainforest Alliance의 진화 "재생농업"
- 제22호 GRI 보고 실무 — Sector Standards와 Content Index
- 제23호 탄소배출국경이 생기자 — EU CBAM과 한국 산업의 과제
- 제24호 격변기의 지속가능 연료 인증: ISCC 제16회 글로벌 컨퍼런스
- 제25호 무엇으로 항해할 것인가 — 대체 해양연료 전환과 인증의 관문
- 제26호 포장에도 여권이 필요하다: EU PPWR과 한국 수출산업의 준비
- 제27호 에너지와 농촌의 지속가능 설계: 햇빛소득마을
- 제28호 규칙은 있고 시장은 없다: RFNBO 인증과 e-연료 시장 창출
- 제29호 감속제가 없는 원자로: 자율 AI의 위험과 한국의 거버넌스 선택



이시도르지속가능연구소

ISIDOR SUSTAINABILITY RESEARCH INSTITUTE

본소 충북 충주시 주덕읍 신덕로 1358 | T.043-845-9792
 청주지소 충북 청주시 봉명로 31 T1타워 1007호 | T.043-900-9792
 연구소 경북 영주시 대학로86번길 14, 2층

www.isidor.kr



ISSN 3092-3549

비매품