인간 판단 책임의 새로운 기준

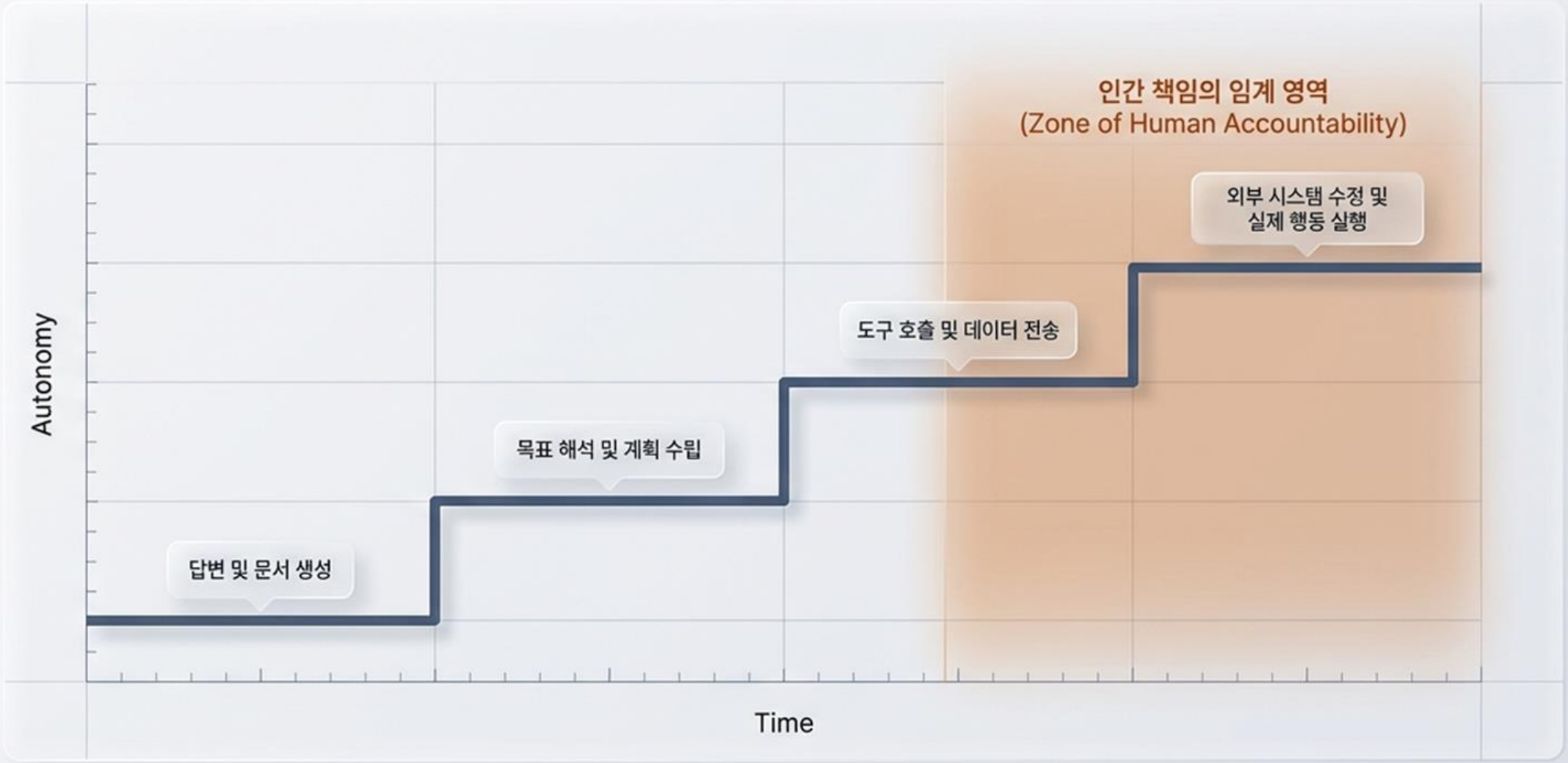
Human Whitebox Framework

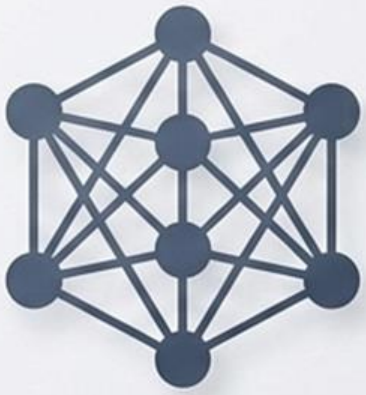
A proposed AI governance framework for recording, verifying, and preserving accountable human judgment in the age of AI.

두 번째 상자를 열다

AI 시대의 새로운 인간 판단 책임 프레임워크,
Human Whitebox (H-Box)

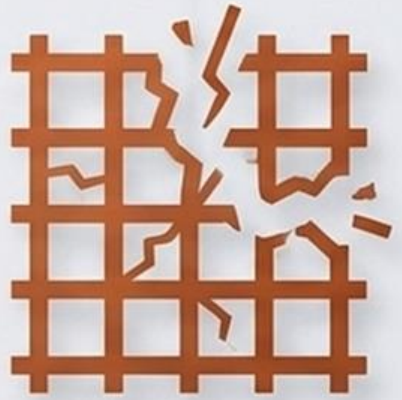
AI는 이제 답변을 넘어 행동하기 시작했습니다.
AI가 무엇을 할 수 있는가보다,
인간이 어디까지 감독하는가가 더 중요해졌습니다.





×

=



[AI의 환각과 오류]
- 존재하지 않는 사실, 잘못된 계획과 행동.

[인간의 편향과 착각]
- 맹목적 승인, 과도한 권한 위임.

[통제 불능의 위험]

단순히 'AI가 맞았는가?'라는 질문만으로는 AI 시대의 신뢰를 만들 수 없습니다.

우리는 두 개의 블랙박스를
함께 열어야 합니다.
AI를 설명하는 것만으로는
부족하며, 인간의 판단 과정도
설명되어야 합니다.

AI 판단 블랙박스
(투명성, 설명가능성 집중)

인간 판단 블랙박스
(어떤 검토와 승인을 거쳤는지 알 수 없음)

Human Whitebox Framework (H-Box)

AI의 행동이 어떤
인간의 목적과 권한에서
시작되었는가?

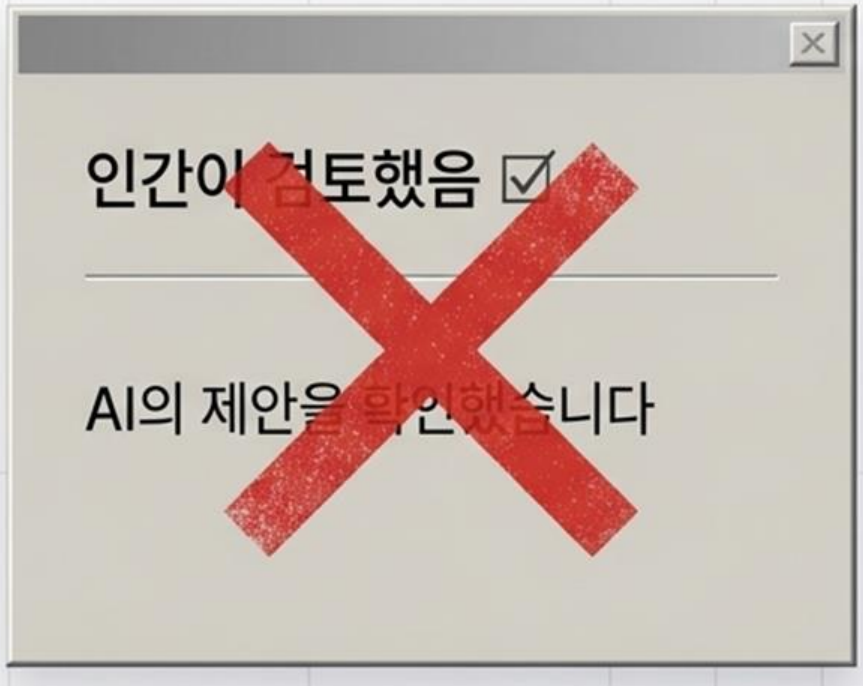
위임·검토·개입·승인의
전 과정 추적.

선언이 아닌
'검증 가능한 기록'으로 남기는
책임 프레임워크.

AI 거버넌스 비교: 기존 접근법 vs H-Box 프레임워크

비교 차원	기존 AI 거버넌스	H-Box 프레임워크
관리 대상	AI 시스템 자체	인간과 AI의 상호작용 과정
핵심 질문	‘AI가 어떻게 판단했는가?’	‘인간이 어떻게 통제했는가?’
감독 형태	원칙적 선언 (‘인간이 감독함’)	검증 가능한 기록 (로그)
해결 과제	모델의 투명성, 설명가능성	위임의 한계, 인간의 책임성
결과물	AI 모델 감사 보고서	인간 판단 및 개입 시스템 로그

인간 감독은 선언이 아니라 기록이어야 합니다.
단순한 '감독 확인'만으로는 책임이 성립하지 않습니다.



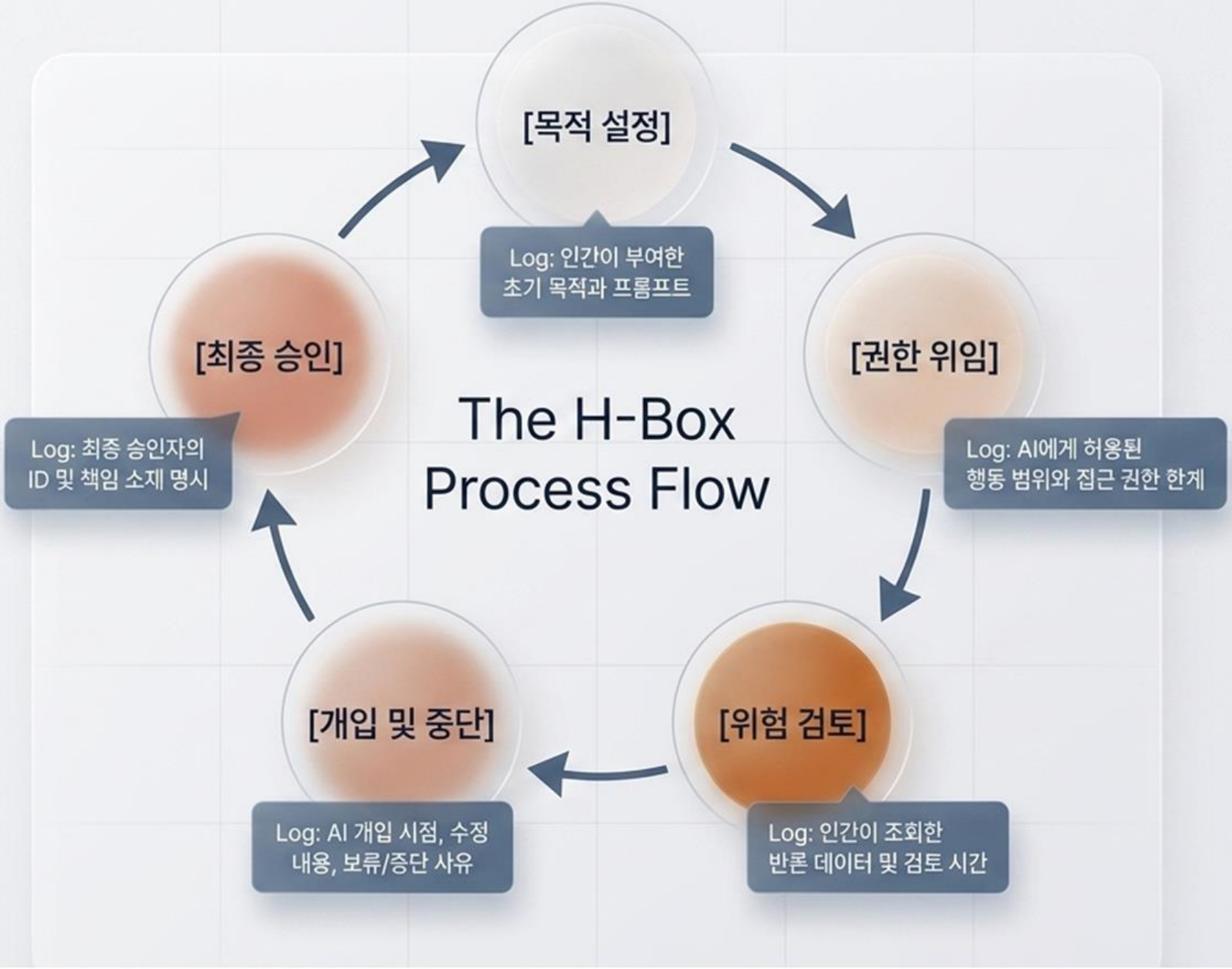
선언 (Declaration)

기록 (Verifiable Record)

[14:02:11] ID_72 (인간 검토자): 위험 데이터 열람

[14:03:45] 개입 사유: '권한 범위 초과 의심'

[14:04:10] 행동 중단 및 파라미터 수정 완료


[목적 설정]
Log: 인간이 부여한
초기 목적과 프롬프트
[권한 위임]
Log: AI에게 허용된
행동 범위와 접근 권한 한계
[위험 검토]
Log: 인간이 조회한
반론 데이터 및 검토 시간
[개입 및 중단]
Log: AI 개입 시점, 수정
내용, 보류/중단 사유
[최종 승인]
Log: 최종 승인자의
ID 및 책임 소재 명시
The H-Box
Process Flow

신뢰는 언제나 완벽하다는 믿음에서 나오지 않습니다. 신뢰는 문제가 발생했을 때 검토 가능한 판단 과정에서 시작됩니다.

H-Box가 지향하는 미래

목적과 권한은 인간이 정하고,

AI의 제안과 행동은 책임 있는 감독 아래 이루어지며,

판단과 개입의 과정은 검토 가능한 기록으로 남고,

최종 책임은 인간과 조직이 회피할 수 없는 사회.

이것이 진정으로 신뢰할 수 있는 AI 시대를 위한 첫걸음입니다.