

ETHOBOT

윤리적 의사결정을 위한 ECD 설계서

게임 과제 · 행동 구조 · 평가 증거 · 로깅 명세

2026.09.05 | 설계 모델 ethobot-ecd-1 | 로그 스키마 1.0.0

1. 평가하려는 것은 '어느 버튼을 골랐는가'보다 '어떤 이유로 판단했는가'이다

Ethobot은 공공 AI의 자원 배분을 다루는 두 에피소드다. 플레이어는 문제를 정의하고 자료를 조사하며, 서로 다른 이해관계자의 관점을 비교한다. 이후 제한된 조건에서 계획을 만들고, 결과와 새 증거를 바탕으로 계획을 재검토하며, 실제 작동하는 이의 신청 경로를 구성한다.

ECD(Evidence-Centered Design)는 이러한 행동에서 무엇을 주장할 수 있는지 먼저 정하고, 그 주장을 뒷받침할 관찰 증거와 과제를 연결하는 설계 방식이다. 이 문서의 C1~C6는 Ethobot을 위한 평가 설계 제안이며, 검증된 심리 척도나 개인의 도덕성을 판정하는 점수가 아니다. [1, 2]

설계 층	Ethobot에서의 역할
역량 모델	문제 인식, 증거 평가, 관점 이해, 가치 조정, 판단 수정, 책임과 구제
과제 모델	냉방 우선순위 사건과 이의 신청 대기열에서 실제로 수행하는 조사·선택·서술·검증
증거 모델	행동과 산출물을 관찰 변수로 기록하고, 대안 설명을 고려하여 해석
기록·해석 체계	원시 이벤트 → ECD 관찰 자료 → 사람의 루브릭 검토 → 검증 후 역량 추론

구현과 검증을 구분한다

- 구현: 두 에피소드의 행동 기록, ECD 연결 정보, 선택 전후 상태, 실패한 시도, 언어·난이도·지원 맥락, 로컬 JSONL 및 관찰 자료 CSV 추출.
- 설계 제안: C1~C6 역량 주장, 질적 루브릭, 평가자 절차, 타당도·신뢰도 검증 계획.
- 아직 주장하지 않는 것: 실제 학습 효과, 평가 점수의 신뢰도, 한영 측정 동등성, 상용 서비스 수준의 검증 완료.

사용 원칙

지원 방식의 선호 자체를 정답으로 점수화하지 않는다. 자동 처리, 사람의 검토, 임시 지원, 보류 모두 사례와 자원 조건을 고려한 설명의 대상이다. 기록을 열었다고 이해했다고 간주하지 않으며, 힌트를 사용했다고 감점하지 않는다. Signal Sort의 점수는 윤리 역량에 합산하지 않는다.

2. Stage 1 — COOL-RANK 과제와 행동 구조

폭염 중 제한된 냉방 시설을 배분하는 상황이다. 이 주민은 주소 이력이 불완전하다는 이유로 순위가 낮아졌다. 플레이어는 주소 공백의 맥락, 집단 신뢰도 0.78과 사례 신뢰도 0.54의 차이, 담당자와 기한이 없는 이의 신청 기록을 조사한다.

과제 묶음	실제 플레이어 행동	남는 증거와 연결
사전 가치·브리핑	가치를 선택하고 미션 단말과 운영 책임자의 설명을 확인	최초 가치·입장. 판단의 출발점이며 그 자체가 숙련도 증거는 아님. C1·C4
문제 정의	결정, 이해관계자, 모르는 점, 가치 충돌을 서술하고 첫 입장 기록	문제 프레임과 초기 이유. 당사자·맥락·가치의 연결을 검토. C1
조사·모델 점검	두 출처 이상의 자료 확인, 투명성 점검 3개, 반례 검증 2개	확인한 자료 ID와 점검 이력. 노출과 이해를 구분. C2
비교 사례	사례의 유사점, 결과에 영향을 주는 차이, 적용 가능성을 서술	단순 복사가 아닌 전이 판단의 근거. 기본 Guided에서 3개. C2·C4
관점 확인	필수 주민·운영 관점을 듣고 충실한 요약 선택, 필요하면 수정	질문, 응답, 요약 선택, 거절·수정. 선택형 인식과 자발적 관점 설명은 별도. C3
숙의·해결안	가치 긴장 선택, 이유 작성, 주장·증거·연결 근거·반론·불확실성 구성	논증과 직접 구성한 지원·검토·감사·책임자·이의 신청·기한. C2·C4
결과·재검토	즉시·지연·반사실 결과를 검토하고 유지·수정·교체·유보 및 이유 기록	전후 계획, 수정 이유, 앞선 논증과의 연결. 변경 유무보다 근거를 평가. C5
책임·성찰	운영 책임자·독립 검토자·24시간 기한 구성, 이의 신청 시험, 성찰	구제 절차의 실행 가능성과 남은 불확실성 설명. C6·C5

지원 수준이 해석을 바꾼다

Stage 1 기본 Guided는 여러 서술 칸에 예시 답안을 제공한다. Supported는 비교·논증 예시를 비우고, Open은 문제 정의·계획 등 더 많은 예시를 비운다. 일부 숙의·영향 설명 예시는 Open에서도 남으므로 ‘Open = 모든 답변이 독립 작성’으로 해석하지 않는다. 지원 설정은 실행 옵션이며 Stage 2 난이도 버튼과 동일한 체계가 아니다.

예시를 그대로 제출한 산출물은 독립적인 추론 능력을 입증하지 않는다. ECD 기록의 difficulty와 scaffold-initialized를 먼저 확인하고, 연구에서는 제공된 예시와 응답을 비교하거나 지원이 없는 별도 전이 과제를 추가한다.

구현 근거: CoolRankVerticalSliceController, ScenarioCommandProcessor, ScenarioLearningState, AgentDrivenConsequenceModel.

3. Stage 2 — 이의 신청 대기열의 행동 구조

자료 5개, 이해관계자 4명, 신청 사례 3개를 비교한다. 이 주민의 주소 공백, 리베라의 완비된 기록, 모건의 공유 주소 표시는 서로 다른 불확실성을 만든다. 자동 처리 1, 사람의 검토 2, 임시 지원 3, 보류 0의 자원을 사용하며 이의 신청용 2단위를 남겨야 한다.

단계	행동 및 진행 조건	해석할 관찰 자료
Briefing	검토 근무 시작	과제 시작·맥락 노출
ProblemFrame	영향받는 집단·빠진 맥락·결정 위험을 20자 이상 서술	C1 후보 증거. 길이는 구조적 입력 조건일 뿐
Investigation	난이도별 3/4/5개 기록 수집	출처 접근 이력. 원문 펼치기와 수집은 다른 행동
Perspectives	난이도별 2/3/4개 관점 확인	선택한 요약·거절·재시도. C3의 제한된 인식 증거
Workbench	세 사례의 처리 결정, 예산 안에서 20자 이상 이유 제출	사례별 선택과 자원 배분 전후 상태. C2·C4
Consequence	두 사람이 같은 주소를 사용한다는 새 사실을 고려하고 이유 제출	최초 결정과 이후 결정의 비교. C5
Redress	책임자·접수 경로·4/12/24시간 목표를 지정하고 경로 시험	온라인 전용 경로의 실패와 전화·대면 경로 재 시험. C6
Debrief	바뀐 판단과 남은 불확실성을 20자 이상 서술	C5 후보 증거. 완료 플래그와 별도로 질적 검토

스토리 난이도	전체 여력 / 계획 상한	필수 기록 / 관점	권장 시간
안내 Guided	12 / 10	3 / 2	시간 압박 없음
보통 Standard	9 / 7	4 / 3	600초
도전 Challenge	6 / 4	5 / 4	420초

자유 조사 모드는 처음부터 자료 5개를 제공하고 Investigation에서 시작한다. 이는 수집 능력이나 자발적 자료 선택의 증거로 사용할 수 없다. 자유 조사는 시간 제한 목표도 없다. 모든 모드에서 시간 경과로 게임이 자동 종료되지는 않는다.

현재 경로 시험은 접수 접근성을 확인하는 제한된 규칙이다. Stage 2에서는 온라인 전용 경로가 실패하고, 전화·대면이 포함된 경로가 통과한다. 이 통과가 모든 현실적 적법성·독립성·서비스 품질을 검증한다는 뜻은 아니다.

구현 근거: `AppealQueueGame.Apply/ValidPlan/Outcome`, `AppealQueueController.Act/Verify`, `AppealMissionGuide`.

4. 역량 모델 — 여섯 가지 검토 가능한 주장

다음 주장은 이 게임의 과제 안에서 드러난 수행을 설명한다. 다른 상황의 행동이나 개인의 윤리적 성품으로 일반화하려면 추가 검증이 필요하다. AI의 신뢰성·공정성·설명 가능성·책임에 관한 내용 구분에는 NIST AI RMF 1.0을 참고하되, 이를 교육 평가 척도로 취급하지 않는다. [3]

ID · 구성개념	주장과 필요한 증거	해석을 제한하는 대안 설명
C1 문제 인식·맥락화	누구에게 어떤 결정이 영향을 주는지, 무엇을 모르고 어떤 가치가 충돌하는지 연결하여 설명	질문 문구·예시 답안 재진술, 글쓰기 능력, 길이 조건 충족
C2 증거 평가·불확실성	자료 출처와 적용 범위를 구분하고, 신뢰도와 자격 판단을 분리하며 선택 이유에 연결	자료 클릭만 수행, 용어 반복, 이미 읽은 자료 목록을 인용으로 오인
C3 관점 이해	상대의 핵심 우려를 과장·축소 없이 나타내고 다른 관점과 비교	두 선택지 중 추측, 반복 시도로 정답 인식, 발언 길이·언어 차이
C4 가치·제약 조정	안전·공정성·속도·자원·이의 제기 간의 긴장을 설명하고 실행 가능한 계획 구성	미리 채워진 계획 제출, 예산 퍼즐만 해결, 특정 개입 선호
C5 근거 기반 재검토	새 증거·결과·대화가 무엇을 바꾸거나 유지하게 했는지 밝히고 불확실성 명시	수정 버튼만 누름, 유지가 무조건 고착이라는 오해, 사후 합리화
C6 책임·실효적 구제	책임자·독립 검토·접근 경로·응답 기한을 연결하고 실패 지점을 보완	게임의 고정 통과 규칙 암기, 실제 조직 역량·접근성 미검증

증거는 강도가 서로 다르다

exposure는 자료 또는 발언을 접한 행동이다. recognition_response는 선택형 요약에 대한 응답이다. action_response는 계획의 구성 요소나 선택을 기록한다. expressed_reasoning은 플레이어가 표현한 이유를 보존하되, 표현의 질이나 독립 작성 여부를 자동 판정하지 않는다.

NPC 발언은 system_output이다. 플레이어가 그 말을 들었다는 사실과 자신의 주장에 사용했다는 사실을 분리한다. ambient-discussion-joined도 대화 노출이지 실제 영향의 입증이 아니다. Stage 1의 social-influence-attribution은 플레이어가 명시한 자기 보고로 해석한다.

사례별 근거 연결

Stage 1의 evidencelds는 명시적으로 기록된 연결이다. Stage 2의 기존 evidencelds는 수집된 자료 전체를 담으므로 evidenceRole을 available_not_cited로 표시한다. 이 목록만으로 ‘플레이어가 해당 자료를 근거로 인용했다’고 추론하지 않는다. 제출한 이유를 사람이 확인해야 한다.

5. 로깅 설계 — 과제에서 증거까지 추적 가능한 구조

기존 ReasoningTraceEvent를 유지하고 ecd 객체를 추가한다. 전송용 계정을 만들거나 서버에 자동 업로드하지 않는다. Windows는 로컬 TelemetryOutput에 JSONL을 기록하고, 웹판은 브라우저 저장 공간과 현재 세션 내보내기를 사용한다. 브라우저 데이터 삭제·시크릿 모드·저장 정책에 따라 지속성이 달라진다.

필드	목적과 해석
sessionId / traceId / sequence	실행 세션, 이어지는 판단 기록, 해당 기록 내 순서. sessionSeconds는 플레이어 수행 시간 전체를 대표하지 않음
ecd.eventId / attemptId	관찰 이벤트의 고유 ID와 이번 실행 구간의 ID. 재개·분기 중복을 구분
taskId / observableId	시나리오/단계와 원본 행동 이름. 번역된 화면 문구 대신 안정적인 식별자 사용
constructIds / evidenceKind	C1~C6 연결 후보와 관찰 종류. 값이 비어 있으면 역량 추론 대상이 아님
outcome / responseCode	관찰·수락·거절·완료 및 규칙 피드백. accepted는 구조·게임 규칙 통과이며 추론 품질 점수가 아님
stateBefore / stateAfter	Stage 2 결정, 최초 결정, 사용 노력, 책임자, 접수 경로, 기한, 시험 결과 등의 전후 상태
context	언어, 모드, 난이도/지원 수준, 요구 자료·관점 수, 버전, 진단 실행 여부
resumedFromSequence	이번 실행이 어느 저장된 판단 순서에서 이어졌는지. 과거 행동을 새 행동으로 다시 세지 않음
evidenceRole / contentTruncated	수집 자료와 인용 자료 구분, 기존 1,200자 기록 한도에 따른 내용 잘림 표시
interpretation	항상 unscored. 자동 도덕성 판정이나 잠재 특성 추정 없음

실패와 재시도가 중요한 이유

Stage 2 Act는 성공뿐 아니라 거절된 명령도 기록한다. 온라인 전용 경로 시험 실패 → 접수 경로 수정 → 재시험 통과를 같은 과제의 행동 연쇄로 읽을 수 있다. 잘못된 관점 요약도 별도 이벤트로 남는다. 단순 성공 횟수보다 수정의 내용과 이유를 확인하는 데 유용하다.

기존 데이터와의 호환

ECD 필드가 없는 과거 로그에 새 역량 정보를 소급해서 붙이지 않는다. 추출기는 이를 legacyWithoutECD로 보고한다. 과거 이벤트를 가져올 때 원래 시간과 식별자를 유지한다. eventId로 복제 이벤트를 제거하고 attemptId·resumedFromSequence로 재개 구간을 구분한다.

6. 행동-증거 사전과 분석용 추출

실제 행동 이름	ECD 연결	기록 해석
problem-frame-defined / appeal-frame	C1 · expressed_reasoning	문제 정의 산출물
evidence-inspection / appeal-evidence	C2 · exposure	자료 접근. 이해 확인 아님
model-transparency-check / red-team-counterexample	C2 · exposure	작성된 점검·반례를 실행한 이력
comparative-case-analysis	C2-C4 · expressed_reasoning	유사·차이·전이 판단
perspective-summary / perspective-revision / appeal-perspective	C3 · recognition_response	요약 선택과 검증 결과
structured-argument / solution-workbench-committed / appeal-commit	C2-C4 · expressed_reasoning	논증·계획·배분 이유
post-consequence-revision / appeal-revise / final-reflection / appeal-reflect	C5 · expressed_reasoning	재검토와 성찰
accountability-component-selected / appeal-route-tested / appeal-test	C6 · action_response	구제 요소 선택과 경로 시험
hint-requested / hint-auto-shown	scaffold	요청한 도움과 자동 안내
locale-changed / checkpoint-resumed	process_context	언어 전환과 저장 재개
agent-response / agent-argument-challenge	system_output	시스템이 제공한 발언·반론
arcade-started / arcade-completed	recreation · 연결 없음	휴식용 미니게임 결과

추출 절차

Tools/export_ecd.py에 JSONL 파일들과 --output 경로를 전달한다. 관찰 자료는 observables.csv, 건수·누락·오류 보고서는 export-summary.json으로 생성된다. 기본 CSV에는 자유 응답을 넣지 않는다. 연구자가 적절한 동의와 보관 절차를 갖춘 경우에만 --include-text를 명시하여 표현 내용을 포함한다.

기술 검증용 자동 플레이 로그는 context.origin=diagnostic으로 구분한다. 학습자 표본과 합치지 않는다. 텍스트를 기본 제외해도 원본 JSONL, 저장 파일, NPC 기억에는 자유 응답이 남을 수 있으므로 단순 가명 ID를 익명화 보장으로 설명하지 않는다.

분석 단위

사건은 단일 eventId, 수행 구간은 attemptId, 이어지는 판단 과정은 traceId이다. 요약 분석은 과제별 증거 프로파일로 시작한다. 자료 수집 횟수·플레이 시간·힌트 횟수를 더해 하나의 윤리 점수를 만들지 않는다. 지도·이동·지원 노출은 대안 설명을 점검하는 맥락 변수다.

7. 루브릭과 평가 절차 — 아직 자동 점수화하지 않는 설계

게임은 아래 점수를 계산하지 않는다. 먼저 평가자들이 실제 산출물을 검토하여 사용 가능한 기준을 정교화하고 일치도를 확인해야 한다. 답변의 길이, 특정 선택, 진행 완료를 질적 평가의 대응으로 쓰지 않는다.

상태/수준	공통 해석
관찰 부족 (NA)	미수행, 기록 누락, 내용 잘림, 예시 그대로 제출 등으로 판단 근거가 부족함. 낮은 역량과 동일하지 않음
1 · 표면적	관련 요소를 이름 붙이지만 사례·행동·근거 사이의 연결이 약함
2 · 연결된 설명	사례의 구체적 증거와 서로 다른 고려사항을 선택이나 계획에 연결
3 · 비판적·수정 가능	대안 설명·반론·불확실성을 다루고, 판단을 유지 또는 바꿀 조건과 실행 가능한 보호 장치를 명시

구성개념	이 게임에서 수준 2~3을 검토할 구체적 질문
C1	주소 이력의 누락을 곧바로 부적격으로 보지 않고, 당사자·결정·가치 충돌을 연결했는가?
C2	집단 신뢰도와 사례 신뢰도를 구분하고, 반례나 자료의 한계를 실제 판단 이유에 사용했는가?
C3	주민의 안전과 운영자의 여력을 모두 충실히 나타내며, 다른 관점을 반론에 반영했는가?
C4	세 사례의 차이와 예산·이의 신청 여력을 고려했는가? 누가 대기 위험을 부담하는지 설명했는가?
C5	공유 주소의 새 정보가 어느 가정과 결정을 바꾸는지 설명했는가? 유지한다면 그 이유와 변경 조건이 있는가?
C6	접근 가능한 접수, 책임자, 독립 검토, 응답 기한을 연결했는가? 통과 버튼 이상의 현실적 한계를 인식했는가?

권장 평가 흐름

- 지원 수준·언어·모드·제공 자료·예시 답안·기록 완전성을 먼저 확인한다.
- 과제별 텍스트와 선택 전후 상태를 묶고, 판단을 뒷받침하는 이벤트 ID를 평가 근거에 남긴다.
- 두 평가자가 독립 평가 후 불일치를 논의한다. 합의 전 원평가를 보존하고 불일치 원인을 기록한다.
- 예비 연구에서 범주 분포, 결측, 평가자 일치도와 근거의 적합성을 검토한다. 통계량의 종류와 표본 수는 연구 질문·설계에 맞춰 사전 결정한다.
- 측정 근거가 확보되기 전에는 총점·순위·인사 또는 고위험 결정을 위해 사용하지 않는다.

8. 타당도 위험과 다음 검증 과제

위험	현재 상태와 대응
예시 답안·지원 효과	Stage 1 지원 수준과 제공 예시를 기록한다. 독립 작성 여부를 확인할 별도 과제와 비교가 필요
언어 동등성	한영 UI와 글꼴을 지원하지만 20/24자 조건의 부담은 언어마다 다르다. 번역·인지 면담·응답 과정 비교가 필요
고정 규칙 학습	관점 요약과 접수 경로 시험은 선택형·규칙형이다. 이유를 설명하는 과제와 새로운 사례에서의 전이를 추가
자연어 결과 판정	Stage 1 결과 모델은 계획 텍스트의 영문·국문 키워드 규칙을 사용한다. 부정·조건문 전체를 이해하는 모델이 아니며 결과 서술을 실제 정책 시뮬레이션 또는 평가 점수로 해석하지 않음
수행 조건 차이	Stage 2 모드·난이도는 자료 수, 여력, 시간, 자동 안내를 바꾼다. 동일 점수 척도로 무조건 합치지 않음
선택 변경의 오해	바꾸는 것은 항상 개선이 아니며, 유지하는 것도 항상 고착이 아니다. 근거와 조건을 평가
맥락 자료와 인용 혼동	available_not_cited를 확인하고, 실제 이유에서 어떤 증거를 사용했는지 사람이 검토
기록 손실·개인정보	잘림 표시·레거시 구분·재개 식별자를 확인한다. 원문에는 개인정보가 포함될 수 있고 연구 동의 관리 UI는 별도 구축 과제

검증 순서

- 기술 검증: 실제 실행에서 JSONL 필드, 순서, 전후 상태, 거절·재시도, 언어 전환, 재개, 내보내기가 일치하는지 확인한다.
- 응답 과정 검증: 대표 사용자의 플레이 관찰과 면담으로 UI·게임 조작이 추론을 가리지 않는지 점검한다.
- 평가 기준 검증: 과제 전문가와 평가자가 산출물을 검토하고, 주장·증거 연결과 루브릭 앵커를 수정한다.
- 경험적 검증: 학습 또는 평가 목적에 맞는 비교 조건, 사전·사후 또는 전이 과제, 언어·지원 수준 통제를 설계한다. 이 단계 전에는 학습 효과를 주장하지 않는다.

근거 문헌

[1] Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus Article: On the Structure of Educational Assessments. Measurement: Interdisciplinary Research and Perspectives, 1(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

[2] Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A Brief Introduction to Evidence-Centered Design. ETS Research Report Series. <https://doi.org/10.1002/j.2333-8504.2003.tb01908.x>

[3] Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. <https://doi.org/10.6028/NIST.AI.100-1>

ECD 문헌은 OpenAlex 서지 검색과 ETS/출판사 원문 페이지에서 확인했다. NIST 문헌은 AI 윤리 내용 영역을 정리하는 데 사용했다. C1~C6의 구체적 과제 연결과 루브릭은 이 프로젝트의 설계 판단이며, 문헌이 Ethobot 자체의 효과를 검증한 것은 아니다.