

RAI Risk Taxonomy 2.0
검수 완료 마스터 릴리스 기술보고서

Human-review recovery and independent audit release

2026년 9월 1일

요약

본 보고서는 네 단계의 작업을 기록한다. 첫째, 798개 기준선 L4 카드에 2차 휴먼 검수 지시(808행 등록부, 검토의견 251건)를 행 단위로 재적용하여 777장의 복구 릴리스를 산출했다. 둘째, 독립 검수 파이프라인 AC-01부터 AC-17까지를 적용하여 629장의 검수 마스터를 산출했다. 셋째, KTSPACE 3차 검수의 629행을 수정 전에 모두 읽고, 비어 있지 않은 검토의견 30건을 재배포·정의 수정 19건, 삭제 6건, 동일 L3 내 범위 일반화 5건으로 해석·적용했다. 넷째, AC-19에서 전문 리뷰어 2인의 Physical AI 독립 검토와 사용자 승인에 따라 도메인 간 재배포 2건과 한국어 용어 교정 1건을 적용했다. 어떤 단계에서도 EM 또는 Hybrid EM을 재실행하지 않았다. **최종 결과는 L4 카드 623장이며 General 494, Agentic 66, Physical 63이다. 47개 실질 L3는 모두 카드를 보유하고, 3개 Others 배정은 0건이며, 50행 L3 마스터는 AC-19 적용 전과 바이트 단위로 동일하다.**

1 복구 원칙

기준 커밋의 유효한 편집과 사용자 승인 통합은 보존했다. 2차 검수 원문의 삭제, 통합, 분리, 재배포, 정의 수정 지시만 source_row_id 계보를 통해 적용했다. 복구 단계에서는 L3 마스터를 보존했으나, 이후 AC-14에서 명시적으로 승인된 G_SYS_PERF 카테고리만 신설하고 마스터 해시를 갱신했다. 휴먼검토의견 251건의 반영 충실성은 별도 검토에서 반영 246, 부분반영 3, 미반영 2로 판정되었고, 각 카드에 검토의견 원문과 반영 결과를 기재했다.

2 적용 결과

조치	검수행 수
재배포	80
정의 수정 후 유지	39
분리	19
통합	14
삭제	14

Table 1: 승인된 166건의 휴먼 검수 조치(복구 단계)

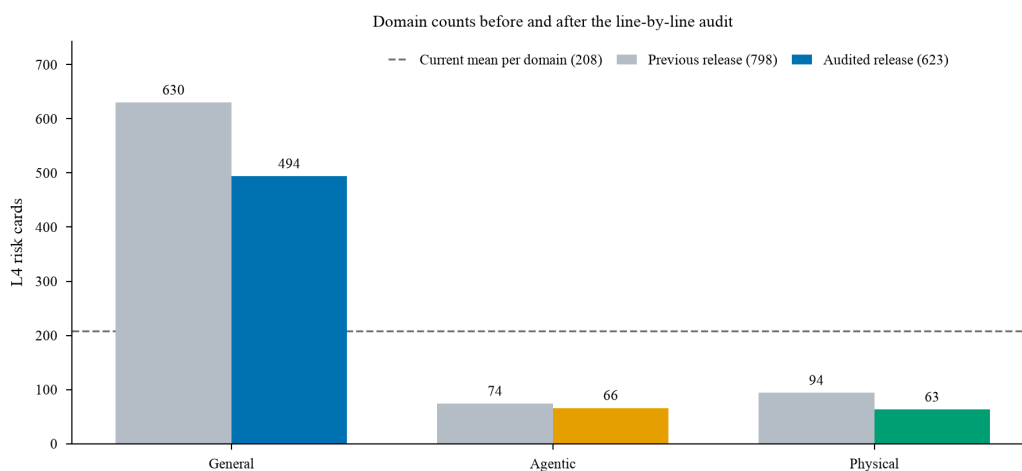


Figure 1: 이전 릴리스와 비교한 현재 도메인별 L4 카드 수

3 감사 시정(AC-01 AC-08)

복구 릴리스에 대한 라인바이라인 검수와 사용자 승인에 따라 다음 시정을 적용했다. 모든 판단적 결정은 리뷰어 2인의 독립 판정 원장과 합의 목록에 기록했고, 합의(교집합) 항목만 적용했다. 2차 휴먼 검수는

시정	내용	카드 수
AC-01	자살 준비·실행 분리 복원(준비 카드 신설, 조장 카드 정의 축소)	777→778
AC-02	8개 대형 L3의 합의 중복 58장 흡수	778→720
AC-03	휴먼 검토의견 원문·반영 결과 기재, 출처 표기 정리	720
AC-04	공개 CSV에서 스코어링·키워드 컬럼 제거(전체 컬럼 원본 별도 보관)	720
AC-05	빈 L3 보충(과도한 거절 3장, 물리적 변조·파괴 3장, 전 카드 출처 링크)	720→725
AC-06	미달 L3 문헌 근거 보강 2장, 거버넌스·조작 계열 합의 16장 흡수	725→711
AC-07	전체 L3 MECE 큐레이션: 흡수 64, 귀속 이관 18, 제거 2	711→645
AC-08	합의 복합 카드 18장 분해(17장 조각 귀속 후 폐기, 위기 대응 실패 카드 신설)	645→629

Table 2: 감사 시정 요약

”부족한 성능·역량”을 다루는 신규 L3의 생성을 명시적으로 요청했다(2건의 신규 L3 요청과 17건의 이관 지시). 이 요청은 L3 마스터 불변 원칙과 충돌하여 보류되어 있었으나, 사용자 승인에 따라 AC-14에서 **성능·신뢰성 실패(Performance and Reliability Failure)** 카테고리를 신설하고 리뷰어 2인 합의로 10장을 이관했다. L3 마스터는 47개 마스터 범주와 3개 Others로 구성되며, 마스터 해시는 이에 따라 갱신되었다. 신설·수정

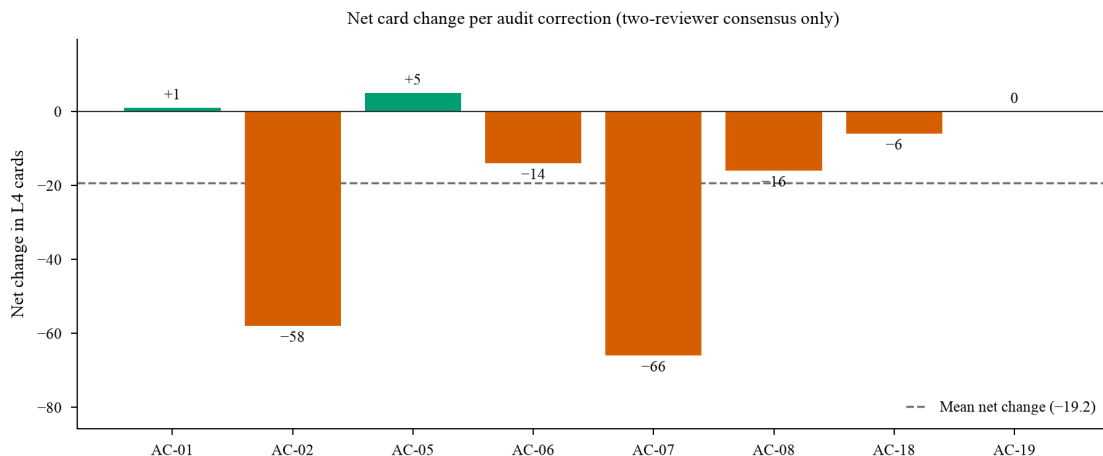


Figure 2: 감사 시정별 L4 카드 순변화

카드는 검증된 문헌 출처를 References로 보유한다(XSTest, OR-Bench, AgentHarm, Cao et al. CCS 2019, NIST SP 800-82r3, NIST SP 800-193, Sclar et al. ICLR 2024, Huang et al. ICLR 2024, Moore et al. FAccT 2025). 자해·자살 계열은 준비, 실행, 조장, 위기 대응 실패의 네 축으로 분리 유지된다.

4 3차 휴먼 검수(AC-18)

3개 원천 표는 General 487행, Agentic 65행, Physical 77행으로 총 629행이다. General 13행과 Physical 17행에 검토의견이 있었고 Agentic의 검토의견 셀은 모두 비어 있었다. 의견이 없는 599행은 내부 전체 컬럼 마스터에서 바이트 단위로 유지했다. 의견이 있는 30행은 수정 전에 현재 카드 의미와 변경되지 않은 L3 마스터에 대조하여 의도를 해석했다. 전문 리뷰어 2인은 명시적 삭제와 대상 카테고리 독립적으로 판정하고 모호한 사례를 별도로 검토했다. 최종 판정자는 G_INT_SELF_006을 G_SYS_CONTEXT로 이동했다. 이는 원문의 맥락 부적합 콘텐츠 기제를 보존하면서 근거 없는 악의적 조작 의도를 추가하지 않기 위한 결정이다. P_INT_SAFETY_007은 통합 지시가 없으므로 범위만 일반화하고 독립 카드로 유지했다. 두 리뷰어는 P_INT_SAFETY_0047 Physical Safety와 충돌하며 G_INT_UNETH에 속한다는 데 합의했다. Physical에서 General로 이동한 모든

3차 검수 조치	카드 수
재배치 및 정의 수정	19
삭제	6
동일 L3 내 범위 일반화	5
통합	0
분할	0
신규 L4 또는 L3	0

Table 3: 3차 휴먼 검수 조치

카드는 물리적 기제와 결과를 명칭·정의에 보존했다. 기존·신규 ID 대응과 삭제 6건은 3차 검수 결정 원장에 기록했다.

5 Physical AI 최소 시정(AC-19)

전문 리뷰어 2인은 Physical AI 카드 65장을 변경하지 않은 L3 마스터와 대조하여 독립적으로 검토했다. 사용자는 가장 확신도가 높은 최소 시정 3건만 승인했다. P_SYS_HARDWARE_001은 물리적 하드웨어의 파손·마모가 아니라 센서 및 분포 변화에 따른 배포 성능 저하를 다루므로 G_SYS_PERF_017로 재식별하여 이관했다. P_INT_TAMPER_001은 물리적 하드웨어 변조가 아니라 적대적 입력 공격인 센서 스푸핑과 신호 주입을 다루므로 G_SYS_SECADV_061로 재식별하여 이관했다. P_SYS_HARDWARE_003은 ID와 계층을 유지하되, cosmicray의 한국어 오역인 “우주선”을 “우주 방사선”으로 교정했다. 삭제, 통합, 분할 또는 신규 카드는 없으며 전체 623장은 유지되고 도메인별 수는 General 494, Agentic 66, Physical 63으로 변경되었다.

6 배포 구성

최종 배포는 L1 CSV 1개, L1/L2/L3 CSV 1개, General, Agentic, Physical L4 CSV 3개로 구성한다. 공개 L4 CSV는 계층 정의(한·영), facet, act-type의 25개 컬럼을 제공한다. 내부 전체 컬럼 원본은 휴먼 검토의견, 반영 결과, 계보, References와 역사적 스코어 필드를 보존하며 웹사이트에는 어떤 스코어도 표시하지 않는다. 검증 폴더에는 629행 3차 검수 결정 원장, 30행 적용 로그, AC-19 시정 원장과 검증 기록, 방법론 문서, 기존·신규 ID 대응표, 삭제 기록, 이전 검수 등록부와 감사 시정 로그를 보존한다.