미국, 유럽, 일본 사례와의 비교를 통한 딥페이크 관련 CITRA 기반 디자인 윤리 가이드라인 제안

(Proposing CITRA-Based Design Ethics Guidelines for Deepfakes: A Comparative Analysis of the United States, European Union, and Japan)

조성배*

(SungBae Jo)

요 약

본 논문은 과학기술정보통신부의 '사람중심 AI 윤리 원칙'을 출발점으로, 딥페이크에 특화된 'CITRA; Consent & Dignity(동의 및 존엄성), Integrity(정직성/맥락보존), Traceability(추적 가능성), Risk & Proportionality(위험 및 비례성), Accountability & Remedy(책임 및 구제책)'디자인 가이드라인을 제안한다. 미국·유럽·일본의 규범·표준·자율 가이드를 비교 참조하고, EU의 AI Act/DSA, 미국의 NIST AI 위험관리 프레임워크 및 생성형 AI 프로파일, 일본 정부 지침, C2PA(Content Credentials), PAI 권고를 근거로 동의, 라벨링/경고, 출처성, 위험등급, 책임/구제 요구를 교차 분석하였다. 그 결과, 제작 - 유통 - 표시 단계의 위험을 CITRA 축으로 재구성하고, 라벨의 가시성·지속성·기계가독성, C2PA 기반 프로비넌스, 고위험 맥락의 비례 통제, 신고 - 삭제 - 복구 - 지원(SLA) 등 실천적 최소 준수선을 도출하였다. 본 연구는 교육·산업·공공 현장에서 즉시 적용 가능한 핵심 체크포인트와 단계적 도입 청사진을 제시한다.

■ 중심어 : 딥페이크 ; 디자인 윤리 ; CITRA 프레임워크 ; 데이터 거버넌스 ; 가이드라인

Abstract

This paper proposes CITRA (Consent & Dignity, Integrity, Traceability, Risk & Proportionality, Accountability & Remedy) design guidelines specialized for deepfakes, starting with the Ministry of Science and ICT's "Human-Centered AI Ethics Principles." By comparing and referencing norms, standards, and voluntary guidelines from the US, Europe, and Japan, and cross-analyzing consent, labeling/warnings, provenance, risk ratings, and accountability/remedy requirements based on the EU AI Act/DSA, NIST AI RMF and Generative AI Profile, Japanese government guidelines, C2PA (Content Credentials), and PAI recommendations, we reframe risks across the production-distribution-labeling stages along the CITRA axes and derive practical minimum compliance standards, including label visibility, persistence, and machine-readability; C2PA-based provenance; proportional controls in high-risk contexts; and service-level agreements (SLAs) for reporting, deletion, restoration, and support. This study presents key checkpoints and a blueprint for phased implementation that can be immediately applied in education, industry, and the public sector.

■ keywords : Deepfakes ; Design ethics ; CITRA framework ; Guideline

Ⅰ. 서 론

1. 연구 배경 및 목적

과학기술정보통신부의 '사람중심 AI 윤리 원칙'은 국내 AI 거버넌스의 상위 규범을 제시하지만, 합성·조작 미디어인 딥페이크에 특화된 디자인 단위의 구체 가이드라인은 여전히 부족하다

* 정회원, 청주대학교 산업디자인학과

접수일자 : 2025년 09월 01일
게재확정일 : 2025년 09월 18일
교신저자 : 조성배 e-mail : josb@cju.ac.kr

[1]. 국제적으로도 유럽연합의 EU AI Act(투명성·표시 의무)와 DSA 선거 무결성 가이드라인(플랫폼 위험경감), 미국 NIST AI 위험관리 프레임워크 1.0 및 생성형 AI 프로파일, 일본의 정부 조달·활용 가이드라인 등은 주로 범용 AI를 대상으로 하며, 실무자가 곧바로 제품·플랫폼의 UI/UX, 메타데이터(출처성), 운영 프로세스로 구현하기에는 세부 설계 기준이 분산되어 있다. 한편 업계 표준·자율 영역의 C2PA(Content Credentials Provenance and Authenticity; 콘텐츠 출처와 신뢰성을 위한 연합)[2]와 PAI(Partnership on AI; AI에 대한 파트너쉽)은 출처성·라벨링·책임 실무의 방향성을 제시하지만[3], 국내 맥락(선거, 성적 이미지·청소년, 저작권·퍼블리시티 등)과의 정합적 적용 기준은 정리되어 있지 않다. 국내 연구·정책 역시 라벨링 - 출처성 - 피해자 구제(SLA)의 통합 대응 필요성을 지속적으로 지적한다[4].

본 논문의 목적은 이러한 공백을 메우기 위해, '사람중심 AI 윤리 원칙'을 토대로 딥페이크에 특화된 디자인 윤리 가이드라인을 제시하는 데 있다. 이를 위해 미국·유럽·일본의 규범·표준·자율 가이드를 비교·참조하여, 제작 - 유통 - 표시 단계에서 적용 가능한 최소 준수선(Baseline)과 운영 로드맵을 제안하고자 한다[5].

그림 1. 과학기술정보통신부 '사람중심 AI 윤리원칙'

2. 연구 범위 및 방법

본 연구는 한국 과학기술정보통신부의 '사람중심 AI 윤리 원칙'을 상위 준거로 삼되, 딥페이크 특유의 위험을 제품·플랫폼 설계 수준에서 다루기 위해 비교 범위를 미국·유럽연합·일본으로 한정하였다. 분석 대상은 첫째, 법·정책 영역의 주요 1차 자료－EU AI Act(합성·조작 콘텐츠 표시의무)와 DSA 선거 무결성 가이드라인, 일본 정부의 생성형 AI 조달·활용 지침－를 포함하고, 둘째, 표준·기술 사양 영역의 C2PA(Content Credentials)를 핵심 준거로 삼았으며, 셋째, 자율 프레임워크로서 NIST AI 위험관리 프레임워크 1.0[6]과 생성형 AI 프로파일[7], 그리고 PAI 합성미디어 책임 관행을 포괄하였다. 적용 단계는 딥페이크의 제작 - 유통 - 표시 전 주기이며, 매체 범위는 텍스트·이미지·오디오·비디오를 모두 포함한다. 특히 선거·공공정보, 청소년·성적 이미지, 저작권·퍼블리시티, 상업·광고 등 고위험 맥락을 우선순위로 다루어 국내 정책·조달·서비스 운영과의 정합성이 높은 최소 준수선을 도출하고자 했다.

방법론은 데스크 리뷰에 기반한 비교 연구로, 선정된 문헌을 먼저 동의, 라벨링/경고, 출처성·추적성, 위험등급별 통제, 책임·구제 등 의무·요구 범주로 정규화하였다. 이어 각 권역의 규범(법·정책) - 표준(기술 사양) - 자율(거버넌스 권고)을 제작 - 유통 - 표시 단계에 매핑하여 공통 분모와 격차를 도출하였고, 이를 한국의 '사람중심' 원칙과 정합성 검토를 거쳐 국내 적용이 가능한 핵심 체크포인트와 단계적 도입 로드맵으로 정리하였다.

Ⅱ. 문제정의와 위험계층화

1. 딥페이크의 정의와 생애주기

딥페이크는 기계학습 기반 합성·변조 기술을 통해 실제와 유사하게 보이도록 생성·편집된 텍스트·이미지·오디오·비디오 등 합성·조작 미디어를 의미한

다. 딥페이크 위험은 제작 - 유통 - 표시의 전 주기에서 상이한 양상으로 표출된다.

제작 단계에서는 동의 부재·권리 침해, 민감한 맥락(미성년자·성적 이미지·생체정보) 노출, 출처 추적 불가가 핵심 위험이다.

유통 단계에서는 업로드 - 재공유 과정에서 라벨·메타데이터가 손실되거나, 추천·광고와 결합해 확산이 가속되는 문제가 발생한다[8].

표시 단계에서는 사용자가 콘텐츠의 합성 여부를 인지하고 검증할 수 있도록 라벨·경고(UI/UX)와 출처성(프로비넌스)이 결합되어야 한다.

표 1. 생성 단계별 위험과 설계 운영 요양

단계	주요 위험	실패 지점(예)	설계·운영 요약
제작	동의·권리 침해, 민감 맥락 노출	모델·데이터 출처 불투명	사전·철회 가능한 동의, 민감 맥락 금지/심사, 생성·편집 로그
유통	라벨·메타데이터 손실, 급속 확산	재인코딩·재공유 시 표식 소실	메타데이터 보존(C2PA), 재공유 시 라벨 유지, 탐지·차단률
표시	사용자 오인, 접근성 미흡	라벨 가시성·지속성 부족	썸네일·재생·공유 화면 상시 라벨, 경고 인터스티셜, 검증 버튼

2. 이해관계자 맵과 고위험 맥락

딥페이크 생태계는 창작자/편집자, 도구·모델 공급자, 권리자/피사체, 플랫폼/유통 채널, 검증·표준 주체(C2PA 등), 이용자/시청자, 피해자·시민사회, 감독기관으로 구성된다. 고위험 맥락은 다음이 대표적이다: 선거·공공정보, 청소년/성적 이미지, 저작권·퍼블리시티/상표, 보이스피싱·사칭 사기, 광고·소비자 기망 등이 대표적이다.[9]

표 2. 이해관계자별 생애주기 책임

이해관계자	제작	유통	표시
창작자/편집자	동의·권리 확인, 변환 기록	업로드 전 라벨·메타데이터 삽입	라벨 문구·대체텍스트 제공
도구·모델 공급자	사용제한·안전가드, 데이터 카드	C2PA/워터마킹 API 제공	라벨·메타데이터 전파 SDK
권리자/피사체	허가·철회 경로	침해 신고·철회	정정·삭제 반영 통지
플랫폼/유통	업로드 필터·탐지, C2PA 검증	재공유 시 라벨·메타데이터 보존	라벨 고정노출, 경고 인터스티셜
검증·표준 주체	표준·테스트셋 제공	검증도구·서버 운영	사용자 검증 UI 가이드
이용자/시청자	출처·라벨 확인	오표시 신고	경고 해석·인지
피해자·시민사회	피해 신고·지원	삭제·차단 요청	재유포 방지 지원
감독기관	가이드·감사 기준	사업자 의무 점검	투명성 보고 검토

3. 위험 계층화(RAG) 개요 및 설계

위험계층화는 영향(Impact)×가능성(Likelihood)×맥락(Context)을 기준으로 RAG 등급을 부여하고, 등급별로 비례적 통제를 누적 적용해야 한다. 등급이 높을수록 라벨→프로비넌스→거버넌스→구제(SLA) 순으로 통제 강도가 강화된다. 고위험군(선거, 청소년/성적 이미지, 생체정보)은 상향 등급(R3 - R4) 기본값을 권고한다[10].

표 3. RAG 등급별 최소 통제 기준

등급	대표 맥락	누적 통제 기준
R1(저)	교육·연구 데모, 명백한 패러디	고정 라벨(문구·아이콘)+캡션, 검증 버튼
R2(중)	상업 캠페인, 인플루언서 협업	R1 + 재공유 라벨 보존, 권리·동의 증빙, 신고 버튼
R3(고)	선거·공인·보건	R2 + 경고 인터스티셜, 사전 심의(법무/윤리), C2PA 의무화
R4(매우 고)	성적 이미지·미성년자·생체정보	업로드 금지/화이트리스트, 자동 차단·신고 트리거, SLA(응답≤1h/삭제≤24h)

III. 국제 비교의 핵심 시사점

1. 규범·표준·자율 가이드 개관

가. 유럽연합(EU)

EU AI Act는 합성·조작 콘텐츠(딥페이크)에 대해 표시(라벨링) 의무를 명시하고 감상·이용을

과도하게 저해하지 않는 “적절한 방식”의 고지를 요구한다. 이는 생성물의 기계가독성·지속성을 갖춘 표시와 사용자 고지의 병행을 시사한다. 또한 DSA 선거 무결성 가이드라인은 선거 기간 플랫폼의 비례적 위험경감 조치(경고 인터스티셜, 조기 완화, 내부 점검)를 제시하여 운영상 통제의 기준점을 제공한다[11].

나. 미국(US)

미국은 포괄적 연방법 대신 미 상무부 국립표준기술연구소(NIST)에서 AI 위험관리 프레임워크[12]와 생성형 AI 프로파일을 통해 출처링·감사 등성·워터마킹/프로비넌스, 라벨링/고지, 모니터 실무 중심의 위험관리를 권고한다. 특히 생성형 AI 프로파일은 생성형 AI 전 주기에 대한 통제 항목을 체계화하여 제품·플랫폼 설계로의 전문화를 촉진한다.

다. 일본(JP)

일본은 소프트로(soft law) 중심 접근을 취한다. 경제산업성의 AI 원칙 구현 가이드라인[13]과 비즈니스를 위한 AI 가이드라인[14]이 기업과 공공조직을 대상으로 책임·투명성·안전성 요구를 통합하고, 2025년 디지털청의 공공부문 생성형 AI 조달·활용 가이드라인은 거버넌스 체계·자문 데스크·조달 연계를 구체화했다. 이는 공공조달을 테코노믹 레버로 활용해 내재화된 거버넌스를 확산시키는 전략으로 볼 수 있다.

표 4. 딥페이크 관련 국제 가이드 범주

권역	법/정책(규범)	자율·운영	딥페이크 특화성
EU	AI Act(Art.50), DSA 선거 가이드	플랫폼 운영 가이드	표시 의무·선거 맥락 규율
US	(연방 포괄법 부재)	NIST AI RMF/GAI Profile	운영 통제·거버넌스 강점
JP	METI – AI Guidelines for Business	공공조달·운영 가이드(디지털청)	조달·운영 연계

2. 딥페이크 관점의 공통점과 한계

미국·EU·일본의 대표적 자료를 딥페이크 관점에서 종합하면, 첫째 표시(라벨링·경고)의 필요성에 사실상의 합의가 존재한다. EU AI Act는 합성·조작 콘텐츠에 대한 명확하고 적절한 표시를 요구하며(Art. 50, Recital 134), 선거 시기에는 DSA 가이드라인이 경고 인터스티셜·조기 완화조치 등 비례적 위험경감 수단을 제시한다. 미국의 경우 연방 포괄법은 부재하지만, NIST AI 위험관리 프레임워크와 생성형 AI 프로파일이 조직 차원의 라벨링·고지·모니터링을 위험관리 체계 안으로 끌어들이는 방향을 제시한다. 일본은 METI 지침과 디지털청 가이드라인을 통해 공공조달·운영 단계에서의 책임·투명성 절차화를 강조한다.

둘째 출처성·추적성에서는 업계 표준인 C2PA가 공통 언어로 기능한다. C2PA 2.2는 생성·편집 이력을 암호학적으로 결박해 재유통 시 검증 가능성을 높이며, 플랫폼·도구 간 상호운용을 전제로 한다. 이는 EU의 표시 의무나 NIST의 위험관리 권고와 결합될 때, 라벨의 가시성·지속성·기계가독성을 뒷받침하는 기술적 기반이 된다.

셋째 위험기반 접근과 운영책임이 공통 축으로 나타난다. EU의 DSA는 비례성 원칙에 따라 선거 등 고위험 맥락에서 통제 강화를 요청하고, NIST 프로파일은 Govern - Map - Measure - Manage의 전주기 거버넌스를 통해 로그·감사·지표 기반 운영을 권고한다. 일본은 조달·자문·교육을 묶은 운영 모델로 거버넌스 내재화를 추구한다. 국내 문헌 역시 피해자 중심의 신속 구제, 증거 보존, 재유포 차단을 통합적으로 요구한다.

동시에 몇 가지 구조적 한계가 반복적으로 관찰된다. 우선, 세 권역 모두 딥페이크 전용 UI/UX 세부 기준(예: 라벨의 위치·지속 노출·접근성, 재공유 시 라벨·메타데이터 보존)에 관한 합의된 설계 지침은 분산되어 있다. 또한 고위험 맥락별 누적 통제의 최소선(예: 선거·청소년/성적 이미지·생체정보)에 대해 법·표준·자율 문서 간 해석의 여지가 커서 현장 적용 시 편차가 발

생할 수 있다. 마지막으로, 신고 - 삭제 - 복구 - 지원(SLA;Service Level Agreement, 서비스 수준 협약)과 같은 시간 기준은 플랫폼 운영 가이드와 형사정책 사이의 연결이 느슨하여, 국내외에서 신속 구제 체계의 정합 설계가 과제로 남아 있다[15].

요약하면, 국제 자료는 표시 - 출처성 - 위험기반 운영의 축에서 상호 보완적 강점을 보이지만, 딥페이크 특화 디자인 단위의 세부 기준과 피해자 구제의 시간표준은 여전히 공백이 존재한다. 본 연구는 이러한 공통분모와 한계를 전제로, 한국의 '사람중심 AI 윤리 원칙'과 정합적인 최소 준수선을 4 - 5장에서 설계 지침과 운영 로드맵으로 제안한다.

3. 한국 적용을 위한 시사점

가. 상위 원칙 ↔ 설계 기준의 연동: 과기정통부 '사람중심 AI 윤리 원칙'을 딥페이크 특화 설계 단위로 번역하기 위해, EU의 표시 의무, NIST의 운영화 지침, 일본의 조달·거버넌스 모델을 핵심 최소선으로 통합한다. 구체적으로는 라벨 가시성·지속성·기계가독성, C2PA 기반 프로비넌스, 고위험 맥락의 비례 통제, SLA 중심의 신속 구제를 국내 기본 체크포인트로 제안한다.

나. 공공조달·표준화 레버리지: 일본처럼 조달 요건(라벨+C2PA, 검증 UI, 로그·감사)과 자문·교육 체계를 묶어 정부·교육기관 도입을 가속한다.

다. 국내 고위험 맥락 대응: 성적 딥페이크와 선거 맥락은 피해자 중심의 신속 구제와 재유포 차단이 핵심임을 국내 연구가 확인한다. 정책·운영에 증거 보존 - 삭제 - 지원의 시간 기준을 명기하고(예: 응답 ≤1h/삭제 ≤24h), 분기 투명성 보고를 권고한다.

Ⅳ. CITRA 기반 디자인 윤리 가이드라인

과학기술정보통신부의 '사람중심 AI 윤리 원칙'을 상위 준거로 삼아, 딥페이크에 특화된 CITRA; Consent & Dignity(동의 및 존엄성), Integrity(정직성/맥락보존), Traceability(추적성/출처성), Risk & Proportionality(위험 및 비례성), Accountability & Remedy(책임 및 구제책) 설계·운영 가이드라인을 제시한다. CITRA는 연구자 제안 프레임이며, 미국·EU·일본의 규범·표준·자율 가이드를 비교 참조해 구성하였다.

1. C－Consent & Dignity(동의·존엄)

초상·음성·이름·상표 등 권리 객체 사용에 대한 사전·철회 가능한 동의를 확보하고, 민감 주체(아동, 성적 맥락, 피해 취약계층)를 선제 보호한다. EU는 합성·조작 콘텐츠의 적절한 표시를 의무화(Art. 50, Recital 134)하여 관람·이용을 저해하지 않는 범위에서 명시적 고지를 요구한다. PAI는 상황·맥락·인물 지위(공인/사인)에 따른 동의·표시 차등을 권고한다. 한국 정책·연구는 성적 이미지 피해에서 피해자 동의·통제권 및 신고 - 삭제 - 지원 흐름의 실효성을 강조한다.

표 5. Consent & Dignity최소 구현 체크리스트

항목	최소 준수선	증빙/로그	실패 징후
권리 확인	인물·브랜드·저작물 사용 동의서 수집(디지털 서명), 민감 주체 화이트리스트/블랙리스트	동의 버전·유효기간·철회 이력	상업 캠페인에서 동의 누락, 철회 후 지속 노출
사용자 고지	합성 표시 문구/아이콘 고정노출(접근성 포함)	UI 캡처, A/B 라벨 인지도 테스트	스크롤/재공유 시 라벨 소실
취약 주체 보호	아동·성적 맥락 업로드 금지 또는 심사	차단 로그, 위반 계정 제재 기록	신고 폭증, 반복 위반 계정 누락

2. I－Integrity(정직성/맥락 보존)

합성임을 명확하게 인지시키고, 맥락 오인(예: 선거·보건 정보)을 최소화한다. 대형 플랫폼에 경고 인터스티셜, 확산 억제, 라벨 유지 등 구체 통제를 제시한다. 조직 내부적으로는 라벨링·오남용 시나리오 점검을 정책화한다.

표 6. Integrity 최소 구현 체크리스트

항목	최소 준수선	KPI/지표	실패 징후
라벨 UX	썸네일·재생 중·공유 화면 상시 가시성	라벨 시인성/지속성 인지도 ≥90%	모바일 재공유 시 라벨 누락
경고/차단	고위험 맥락 경고 인터스티셜·속보 확산 억제	경고 노출 대비 이탈률, 확산 속도 감소	선거 주간 확산 속도 급증
휴먼 검토	고등급 콘텐츠 사전 검토(법무·윤리)	검토 리드타임, 반려율	검토 누락·야간 시간대 사고

3. T－Traceability(추적성/출처성)

C2PA로 출처·편집 이력·버전을 암호학적으로 묶고, 재유통 시에도 라벨+메타데이터가 보존·검증되게 한다. C2PA는 Assertions(이력), Claim(서명 묶음), 검증 워크플로를 제공하며, 표준 가이던스는 제작 - 편집 - 배포 툴 간의 상호운용을 다룬다. 프로비넌스/워터마킹을 통제 옵션으로 제시한다.

표 7. Traceability 최소 구현 체크리스트

항목	최소 준수선	구현 포인트	검증/모니터링
메타데이터	C2PA Assertions로 생성·편집·도구 이력 기록	툴 전환 시 손실 없는 전파	무작위 샘플 C2PA 검증 통과율 ≥ 99%
소프트 바인딩	원본·파생 자산 고유 식별자 연결	XMP DocumentID/InstanceID 연동	파생-원본 역추적 성공률
공개 검증	사용자 검증 보기(검증 결과/체인)	UI에 “출처 보기” 제공	검증 실패·메타데이터 제거 탐지 알림

4. R－Risk & Proportionality(위험 및 비례성)

맥락·영향·가능성에 따라 RAG 등급을 부여하고 비례적 통제(라벨→프로비넌스→거버넌스→구제)를 누적 적용한다. Govern - Map - Measure - Manage 기능군과 위험 목록·관리 행동을 제공하고, 선거·아동·생체정보 등 고위험 맥락에서 강화 의무를 시사한다. 한국에서는 성적 이미지·선거 영역의 고등급 기본값과 신속 구제를 강조한다. 이를 바탕으로 최소 구현 체크리스트를 작성하면 아래와 같다.

표 8. Risk & Proportionality 최소 구현 체크리스트

등급	대표 맥락	누적 통제(최소선)
R1(저)	교육·연구 데모, 명백한 패러디	고정 라벨(문구·아이콘)+캡션, 검증 버튼
R2(중)	상업 캠페인, 인플루언서 협업	R1 + 재공유 라벨 보존, 권리·동의 증빙, 신고 버튼
R3(고)	선거·공인·보건	R2 + 경고 인터스티셜, 사전 심의(법무/윤리), C2PA 의무화
R4 (매우 높음)	성적 이미지·미성년자·생체정보	업로드 금지/화이트리스트, 자동 차단·신고 트리거, SLA(응답≤1h/삭제≤24h)

5. A－Accountability & Remedy(책임·구제)

책임 주체(제작자·도구공급자·플랫폼·감독기관) 간 책임 귀속을 명확히 하고, 피해 발생 시 신고 - 삭제 - 복구 - 지원의 시간기준(SLA)과 증빙 로그를 운영한다. EU는 플랫폼 보고·감사와 위반 시 제재 구조를, 일본은 정부 조달·활용 가이드라인을 통해 내부 통제(심사·감사·교육)를 제도화한다. 한국 연구는 성적 촬영물/딥페이크 성범죄의 양형·수사 개선과 피해자 보호 흐름 강화를 제안한다.

표 9. Accountability & Remedy 최소 구현 체크리스트

항목	최소 준수선	운영 지표	참고
책임 매트릭스	이해관계자별 RACI(제작자·도구·플랫폼·감독)	분쟁 시 책임 귀속 평균 결정시간	내부 정책/약관 반영
신고·삭제	원클릭 신고→평가→삭제/차단→복구	신고→ 응답 ≤1h, 삭제 ≤24h(고위험)	국내 보고서·판례 분석
피해 지원	리퍼럴(상담·법률·플랫폼 지원)	지원 연결율/재노출 방지율	여성정책연구원·입법조사처 제언 반영

Ⅴ. 체크리스트와 운영 로드맵

CITRA 원칙을 제품·플랫폼·기관 운영에 곧바로 적용할 수 있도록 핵심 체크포인트와 단계적 도입 로드맵으로 간명화하여 제안하고자 한다.

1. 핵심 체크포인트

CITRA 원칙에 입각한 핵심체크포인트를 아래 표에 근거해 최소한의 준수선을 설정한다.

CITRA 원칙	무엇 (What)	왜 (Why)	최소 준수선 (Baseline)	책임 주체
C (동의·존엄)	사전·철회 동의	권리·인권 보장	인물·음성·상표 등 권리 객체 전자 동의·철회 기록	제작자·플랫폼
I (라벨/경고)	합성 표시 및 맥락 경고	사용자 오인 방지·투명성	썸네일·재생·공유 화면 상시 라벨, 선거·청소년/성적 이미지 등 고위험은 열람 전 경고	플랫폼·서비스
T (출처성/ 추적성)	C2PA 기반 프로비넌스	재유통 시 검증 가능	C2PA Manifest 삽입·검증, XMP ID 등으로 원본-파생 추적, 출처 보기 버튼	도구·플랫폼
R (위험·비례)	RAG 등급 기반 통제	과잉·과소 규제 방지	R1-R4 등급 부여, R3-R4는 사전 심의·경고·업로드 제한 누적	정책·플랫폼
A (책임·구제)	신고-삭제-복구-지원	피해 최소화	SLA: 응답 ≤1h, 삭제/차단 ≤24h(고위험), 증거 보존·이의·복구 절차	플랫폼·기관

2. 단계적 도입 로드맵

CITRA 원칙을 반영한 6단계의 실행 로드맵을 아래 표와 같이 설정하여 실행하고 체크한다.

단계	핵심 작업	산출물	성공 기준(예)
1단계: 준비	책임자 지정, 위험 맵·RAG 초안	거버넌스 메모	조직 커버리지 100%
2단계: 파일럿-1	라벨 UX A/B, C2PA PoC	라벨 가이드, 샘플 검증 리포트	라벨 인지도 ≥90%
3단계: 파일럿-2	재공유 라벨 보존, 검증 UI	기능 출시 노트	보존율·검증율 ≥99%
4단계: 확장	고위험(R3-R4) 심의·경고, SLA 베타	SOP·심의 로그	응답/삭제 ≤ 1h/≤24h
5단계: 정착	투명성 보고·외부 감사	분기 레포트	재업로드율 분기 감소
6단계: 내재화	조달·교육 연계, 표준화 참여	조달 체크리스트	조달 반영·교육 이수율

Ⅵ. 결론

본 연구는 과학기술정보통신부의 사람중심 AI 윤리 원칙을 상위 준거로 삼아, 딥페이크 전 주기(제작 - 유통 - 표시)의 위험을 CITRA라는 기준을 설정하여, 미국·EU·일본 자료를 비교 참조하여 실천적 최소 준수선을 제안하였다. 핵심 메시지는 다음과 같다. 첫째, 라벨/경고의 가시성·지속성·기계가독성 확보가 출발점이다. 둘째, C2PA를 중심으로 한 출처성/추적성은 라벨을 실질화하는 기술적 토대다.

셋째, 위험등급(RAG)에 따른 비례 통제는 선거·청소년/성적 이미지·생체정보 등 고위험 맥락에서 누적적으로 강화되어야 한다. 넷째, 책임·구제－신고→응답 ≤1시간, 삭제/차단 ≤24시간－와 증거 보존·이의·복구는 피해자 중심 운영의 최소선이다.

정책·조달·플랫폼 설계 관점에서의 기여는 첫째, 사람중심 원칙과 정합적인 핵심 체크포인트(라벨/경고 - C2PA - SLA)의 묶음 제시이다. 둘째, 공공·교육 부문 조달 요건화와 투명성 보고를 포함한 도입 청사진 제공이다. 셋째, 거버넌스(로그·감사, 교육)로 이어지는 운영 루프의 구체화에 있다. 이는 범용 AI 중심의 국제 지침을 딥페이크 특화 설계 기준으로 정리해 한국 맥락에서 바로 적용 가능하도록 했다는 점에서 실무적 가치를 갖는다.

한계로는, 본 연구가 데스크 리뷰에 기반하여 실증실험 설계·결과가 제외했다는 점과 각 세부 디자인 영역별 로드맵을 제시하지 못했다는 점이다, 그리고 AI 법·표준의 빈번한 개정으로 인해 권고 기준의 주기적 갱신이 필요하다는 점이 있다. 향후 연구는 라벨 문구·위치·지속성·접근성의 현장 테스트와 플랫폼·메신저·편집툴 간 메타데이터 보존/변조 내구성 평가, 그리고 SLA - 재유포율 - 피해 회복의 상관 분석 등 국제 표준 정합성 검증으로 이어져야 한다.

결론적으로, 본 논문이 제시한 CITRA 기반 최소 준수선과 도입 로드맵은 한국의 사람중심 AI

윤리원칙을 딥페이크 설계·운영으로 연결하는 실행 가능한 기준선이다. 학계·정부·산업·교육 부문이 이를 공통 언어로 활용한다면, 안전과 창작·교육의 균형을 유지하면서도 책임성과 신뢰성을 갖춘 합성미디어 생태계를 조성할 수 있을 것이다.

REFERENCES

[1] 과학기술정보통신부, “사람이 중심이 되는 인공지능(AI) 윤리기준,” *보도자료/정책자료*, 2020년 12월

[2] Coalition for Content Provenance and Authenticity (C2PA), “C2PA Technical Specification, ver. 2.2”. https://spec.c2pa.org/specifications/ (accessed Aug., 4, 2025).

[3] Partnership on AI, “Responsible Practices for Synthetic Media: A Framework for Collective Action”, https://syntheticmedia.partnershiponai.org/ (accessed Aug., 5, 2025).

[4] 김봉섭, “2023년 사이버폭력 실태조사 결과와 의미”, *지능정보윤리 이슈리포트*, 제5권, 제1호, 23-34쪽, 2024년

[5] Sung Bae Jo, Jae Ick Lee, "Proposal of GUI Guidelines to Improve the Usability of Mobile Healthcare for New Silver Generation," *Smart Media Journal*, Vol. 7, No. 2, pp60-70, 2018

[6] National Institute of Standards and Technology (NIST), “Artificial Intelligence Risk Management Framework (AI RMF 1.0)”, https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10 (accessed Aug., 4, 2025).

[7] National Institute of Standards and Technology (NIST), “Artificial Intelligence Risk Management Framework: Generative AI Profile” https://doi.org/10.6028/NIST.AI.600-1 (accessed Aug., 5, 2025).

[8] European Commission, “Guidelines under the Digital Services Act to Mitigate Systemic Risks Online for Elections” https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1707 (accessed Aug., 6, 2025).

[9] 정준화, “인공지능의 내재적 위험과 입법·정책 과제,” *국회입법조사처 NARS 입법·정책,* 제162호, 11-13쪽, 2024년 12월

[10] 강준모, “딥페이크 관련 국내외 규제 동향 분석,” *KISDI AI Outlook*, 제18권, 8-13쪽, 2024년 9월

[11] European Commission, “Commission Publishes Guidelines under the DSA for the Mitigation of Systemic Risks Online for Elections,” *Press Release*, Mar. 26, 2024.

[12] ByungRae Cha, MyeongSoo Choi, EunJu Kang, Sun Park, JongWon Kim, "Trends of SOC & SIEM Technology for Cybersecurity," *Smart Media Journal*, Vol. 6, No. 4, pp.41-49, 2017

[13] Digital Agency (Government of Japan), “The Guideline for Japanese Governments’ Procurements and Use of Generative AI”, https://www.digital.go.jp/en/news/3579c42d-b11c-4756-b66e-3d3e35175623 (accessed Aug. 9, 2025).

[14] Ministry of Economy, Trade and Industry (Japan), “AI Guidelines for Business Ver. 1.0”, https://www.meti.go.jp/english/press/2024/0419_002.html (accessed Aug. 9, 2025).

[15] Dasol Kim, Sicheon You,"Concept and characteristics of safety information design that reflects human characteristics," *Smart Media Journal*, Vol. 13, No.8, pp.79-85, Aug. 2024

저 자 소 개

조성배(정회원)

2000년 건국대학교 산업디자인학과 학사

2004년 Designskolen Kolding, Interactive Media & Industrial Design, 석사

2012년 명지대학교 디자인학과 박사 수료

2012년~현재 청주대학교 산업디자인학과 부교수

<주관심분야 : 산업디자인, UX디자인, 서비스디자인>