『**2025** 제**3**회 **KISIA** 정보보호 개발자 해커톤』

개발기획서

팀명	**AEGIS**
프로젝트명	**AI** 기반 워터마크를 이용한 위변조**(**딥페이크**)** 탐지 및 검증 서비스
프로젝트 소개	본 서비스는 AI 기반 워터마킹 모델을 활용해 미디어에 워터마크를 삽입함. 이후 의심스러운 미디어가 주어지면, 워터마크를 추출하여 위변조된 영역을 탐지 및 위변조 여부 검증
팀 소개 및 팀원별 역할	팀장 - 최준혁 (총괄 기획 및 개발 보조) 팀원 - 강대현 (서비스 시스템 아키텍쳐 설계 및 구현) 팀원 - 조혜원 (사업환경 분석 및 마케팅 전략 제시) 팀원 - 황대연 (AI 모델 구현)

① 추진 배경 및 필요성

.1. 추진 배경**:** 생성형 **AI**가 촉발한 디지털 신뢰의 위기

AI 기술의 비약적인 발전은 사회 전반에 혁신을 가져왔지만, 동시에 생성형 AI(AIGC)를 악용한 디지털 콘텐츠 위변조라는 심각한 부작용을 낳음. 이는 단순한 기술적 문제를 넘어, 사회의 신뢰 기반을 위협하는 현실적인 위기로 다가옴.

- 가짜 뉴스와 허위 정보의 확산
- 정교하게 조작된 이미지는 여론을 왜곡하고 사회적 갈등을 증폭시킴.
- 과학기술정보통신부 주관 2024 대국민 설문조사에서 응답자의 94.5%가 심각한 사회 문제로 인식, 41.9%는 조작 판별에 어려움을 느낌

과학기술정보통신부 보도 자료 당시 대한민국! 새로운 국민의 나라

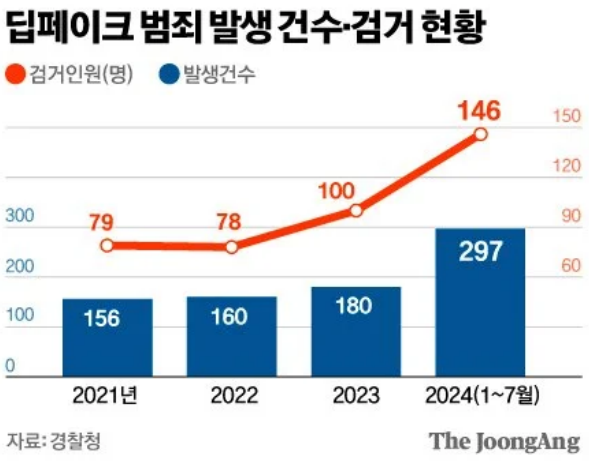
보도시점 2024. 12. 9.(월) 12:00 (2024. 12. 10.(화) 조간) 배포 2024. 12. 9.(월) 10:00

응답자 41.9%, "이미지 영상 조작(딥페이크) 가짜뉴스 판별 못해"

- 대국민 설문조사에서 응답자 39%가 "이미지 영상 조작(딥페이크) 가짜뉴스를 접해봐", 응답자들은 이미지 영상 조작(딥페이크) 가짜뉴스에 대해 강력한 입법 및 정책, 처벌 등 강조

⇒ 언론의 신뢰도 저하를 넘어, 사회 시스템 전반에 대한 불신 초래

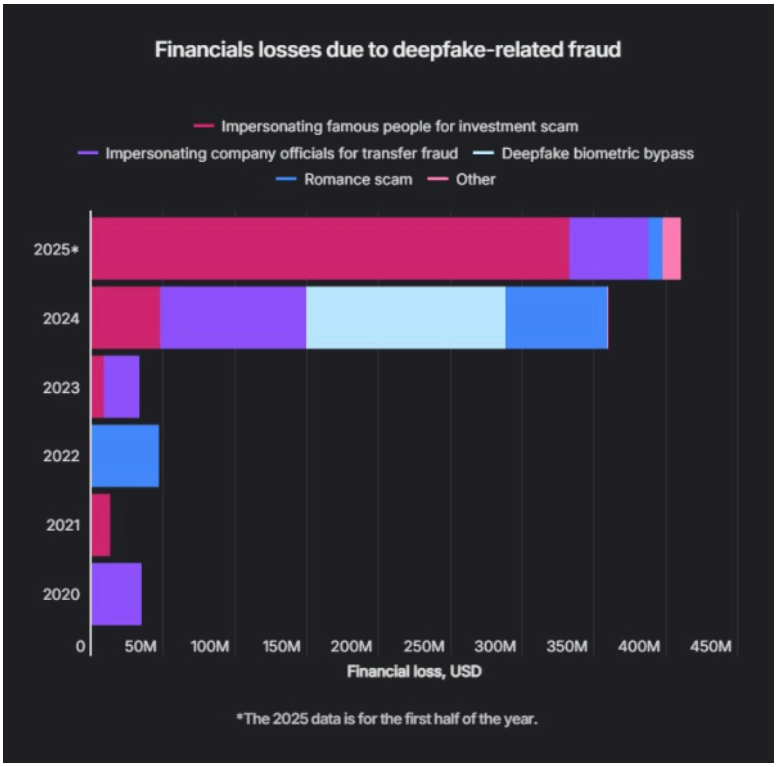
- 명예훼손 및 인권 침해
- 2024 방송통신심의위원회 통계에 따르면, '합성·편집'(딥페이크 포함) 관련 디지털 성범죄 심의는 전년 대비 3.3배 가까이 급증함.
- 3년간 방심위 심위는 5배, 피해자 지원은 약 7배, 경찰 신고는 6배 이상 증가하며, 딥페이크 범죄가 확산되고 있음.



-

⇒ AI 기술 악용이 개인의 삶에 돌이킬 수 없는 피해 제공

- 경제적 손실 및 사회적 비용 증가
- 실제 딥페이크 기반 사기로 인한 전 세계 피해액은 2024년 상반기에만 약 5,600억 원(4억 1천만 달러)에 달하며, 이는 이미 작년 한 해 피해액 초과.
- 딥페이크 사기는 유명인·임원 사칭 등의 수법으로 금융권을 중심으로 막대한 금전적 손실을 야기할 뿐만 아니라, 기업의 신뢰도 하락과 같은 심각한 평판 리스크로 이어짐.



⇒ AI로 조작된 미디어는 개인과 기업의 자산을 직접적으로 위협, 관련 범죄 대응을 위한 사회적 비용 급증

2. 기존 대응책의 명백한 한계

이러한 위협에 대응하기 위한 기존 기술들은 빠르게 발전하는 AI 앞에 손쉽게 무력화되거나, 법적 분쟁에서 실효성 있는 증거를 제시하지 못하는 명백한 한계를 가짐.

- 한계 ① - 손쉬운 무력화: 현재 널리 쓰이는 로고 형태의 가시성 워터마크는 빠르게 발전하는 생성형 AI를 통해 일반인도 손쉽게 제거할 수 있어, 가장 기초적인 보호 수단으로서의 의미를 상실함
- 한계 ② - 법적 증명 능력의 부재: 이를 보완하기 위한 기존 비가시성 워터마크 역시, '변조 여부'라는 단편적인 정보만 제공할 뿐, 법적 책임을 묻기 위한 핵심 요소들을 증명하지 못함.

-

- 변조의 **‘의도성’** 입증 실패: 선의의 편집과 악의적 조작을 구분하지 못해, 법적 책임의 핵심인 ‘고의성’ 입증에 실패함. 실제 판례로, 자신의 사진 워터마크를 무단으로 삭제했다며 제기된 손해배상 소송에서 법원은 피고의 고의성이 입증되지 못해 원고의 청구를 기각한 바가 있음. 이는 '워터마크가 사라졌다'는 사실만으로는 악의적 제거 행위를 입증하기에 불충분함을 보여주는 명백한 사례임.
- 구체적 피해 내용 증명 불가: 초상권, 명예훼손, 동일성유지권 등 다양한 권리 침해를 주장하려면 ‘어떻게’ 변조되었는지 구체적으로 보여줘야함. 단순 워터마크는 변조 여부만 알려줄 뿐, 공격자의 의도를 파악하거나 조작의 심각성을 평가하는데 필요한 위치 특정 정보를 제공하지 못함.
- 손해 규모 산정의 어려움: 피해 규모는 변조의 범위와 심각성에 따라 결정됨. 하지만 변조된 영역과 정도를 특정할 수 없으면 객관적인 피해액 산정이 불가능해 피해자가 정당한 배상을 받기 어려움.

3. 본 서비스의 제안**:**

본 서비스는 딥러닝 기반 워터마킹 모델 EditGuard를 활용하여, 기존 기술의 한계를 극복하고 단순 탐지를 넘어 ‘결정적 증거’를 제공하는 새로운 정보보안 서비스를 개발하고자 함.

기능	본 서비스	기존 비가시성 워터마킹 모듈 (M사)	기존 가시성 워터마킹 모듈 (자사 로고)
신종 AI 변조 대응	O (즉시 대응 가능 - Zero shot)	△ (신규 공격에 대한 모델 재학습 필요)	X (생성형 AI로 제거 용이)
변조 위치 특정	O (95% 이상 정밀도)	X (기능 부재)	X (기능 부재)
워터마크 생존력	O (딥러닝 기반 높은 강인성)	△ (자체 알고리즘 기반, 노이즈에 취약)	△ (원본 훼손 및 시각적 방해)
업그레이드 유연성	O (모듈식 설계)	△ (일체형 구조)	X (기능 고정)

- 독창성 ①**:** 기술적 지속가능성 및 미래 위협 대응력

-

- **AI** 위협에 대한 포괄적인 방어 능력: 딥러닝 기반의 워터마킹 모듈은 압축, 왜곡 등 일반적인 조작에 대한 본질적인 강인성을 가짐. 또한, 특정 변조 유형을 학습할 필요 없는 제로샷(Zero-shot) 탐지 방식을 통해, 끊임없이 등장하는 신종 AI 편집 기술에도 모델 재학습 없이 즉시 대응 가능함.
- 지속가능한 기술 경쟁력을 위한 모듈화 구조: AI 기술은 창과 방패의 끊임없는 경쟁과 같아 보안 제품이 쉽게 뒤처질 수 있음. 본 서비스는 레고처럼 AI 서버를 교체할 수 있는 모듈화로 설계되어서 업그레이드가 유연함.

  ⇒ 현재와 미래의 위협에 모두 대응하는 포괄적인 방어 능력과, 장기적인 기술 경쟁에서 지속적인 우위를 점할 수 있는 경제적 효율성을 동시에 확보

- 독창성 ②**:** 단순 탐지를 넘어선 **'**결정적 증거**'** 생성
- 위변조 영역의 정밀 시각화를 통한 명백한 증거 능력: 본 서비스는 단순히 미디어의 위변조 여부를 판단하는 것을 넘어, 조작된 영역을 마스크(Mask) 형태로 정확히 추출하고 시각화함. 이는 조작의 범위와 내용을 부인할 수 없는 법적 증거로 제공하여, 기술적 분석과 법적 대응의 용이성을 크게 향상시킴.
- 사전 방어 기반의 신속한 **'**무결성 증명**':** 콘텐츠 생성 및 배포 시점에 워터마크를 사전 적용함으로써, 위변조 논란 발생 시 제3의 기관에 분석을 의뢰하고 기다릴 필요 없이 즉각적으로 원본을 증명하고 무결성 검증 공표 가능

  ⇒ 기업 및 공인이 평판 하락을 막고, 가짜 뉴스와 허위 정보가 확산되는 것을 미연에 방지하는 가장 효과적인 수단

② 주요 기능 및 개발환경

-

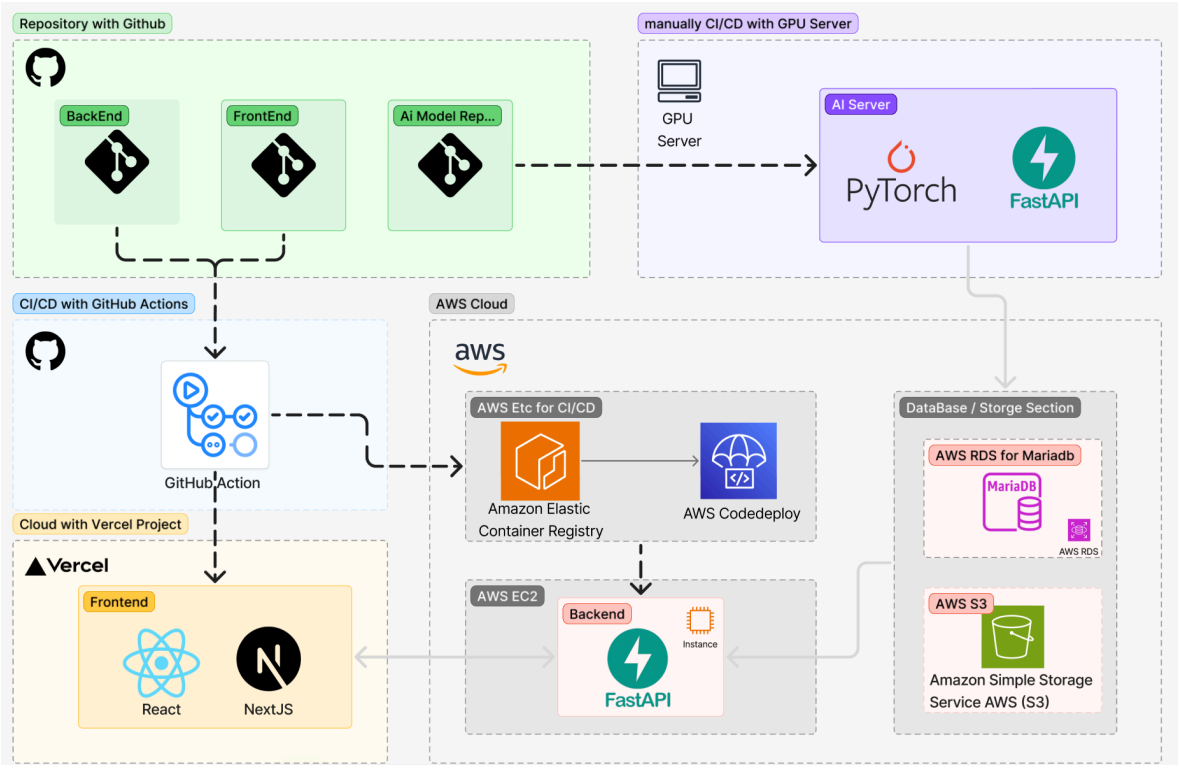
기능을 구현하기 위해 서술한 개발환경이 적절한가? 등 작성 *ex)*개발 플랫폼, 언어, *API* 등 작성

1. 주요기능

본 서비스는 사용자가 안심하고 미디어를 공유 및 활용할 수 있도록 다음의 기능을 제공함.

- 기능 ①: 디지털 원본 증명 및 보호
- <입력> 사용자가 보호할 원본 이미지를 웹 인터페이스에 업로드
- <처리> Dual-Watermark Encoder 기반 보이지 않는 이중 워터마크 삽입
  - 2D 위치 워터마크 (무결성 증명용)
  - 1D 저작권 워터마크 (소유권 증명용)
- <출력> 다운로드가 가능한 워터마크가 삽입된 보호 이미지 제공
- 기능 ②: **AI** 기반 위변조 검증 및 시각화
- <입력> 사용자가 검증할 의심 이미지를 웹 인터페이스에 업로드
- <처리> 워터마크 추출 및 원본과 비교 분석을 통한 위변조 영역 탐지
- <출력> '무결성 검증 보고서' 형태로 결과 제공
  - 위변조 여부 판정: Yes / No
  - 위변조 영역 시각화: 조작된 영역을 시각적으로 표시하는 이진 마스크(Binary Mask)
  - 추출된 원본 정보: 이미지에 인코딩된 저작권 정보

2. 시스템 아키텍처 및 개발 환경



-

본 서비스는 안정성과 확장성을 최우선으로 고려하여, Frontend는 Vercel에, Backend에 배포함. AI 모델은 원활한 AI Serving이 가능하도록 환경설정한 서버에 배포함.

모든 백엔드 서비스는 Docker 컨테이너로 관리되며, Github Actions를 통해 CI/CD 파이프라인을 자동화하여 개발 및 배포 효율성을 극대화함.

- **Frontend Server(Vercel)**
  - 기술 스택: React, NextJs
  - 배포: Vercel을 통해 글로벌 사용자에게 빠르고 안정적인 UI/UX 제공
- **Backend Server(AWS EC2)**
  - 설계: Docker Compose를 활용하여 각 기능을 독립된 컨테이너로 운영하여 안정성 및 확장성 확보
  - 기술스택
    - NGINX: 리버스 프록시로써 트래픽 분산 및 SSL 처리를 담당
    - Backend (FastAPI): Python FastAPI 기반으로 RESTful API를 제공하며, Frontend Server와 AI Server 사이에서 요청을 처리함.
  - **CI/CD (**자동화 파이프라인**)**
    - 프로세스: Github Actions가 코드 변경을 감지하여 Docker 이미지를 빌드하고 AWS ECR에 푸시하면, AWS CodeDeploy가 이를 EC2 인스턴스에 자동으로 배포
- **AI Server(External Server)**
  - 설계: AI Server은 Backend 요청에 의하여 받은 데이터를 AI모델을 통해 이미지의 디지털 원본 증명 및 보호 및 **AI** 기반 위변조 검증 및 시각화을 수행하고 결과를 Backend Server에 반환함.
  - 기술스택
    - PyTorch: 입력된 이미지를 텐서로 변환한 뒤, PyTorch로 구현된 EditGuard 모델을 통해 이미지 위변조 탐지 및 증명 기능을 수행
    - FastAPI: Python 기반의 FastAPI 프레임워크를 활용하여 RESTful API를 제공하며, Backend Server로부터의 요청을 수신하고 결과를 반환하는 역할
- Database
  - 사용 서비스: AWS RDS for MariaDB
    - 구성 목적: 서버와 데이터베이스를 분리하여 확장성과 유지보수성을 향상시킴 AWS RDS를 활용해 MariaDB를 관리함으로써 인프라 관리 부담을 줄이고, 자동 백업, 장애 복구 등의 기능을 통해 안정적인 데이터 관리를 구현
    - **Backend Server** 또는 **AI Server**에서 생성된 로그**,** 분석 결과**,** 사용자 요청 데이터 등을 저장하는 주요 저장소 역할
- Storage
  - 사용 서비스: Amazon S3 (Simple Storage Service)

-

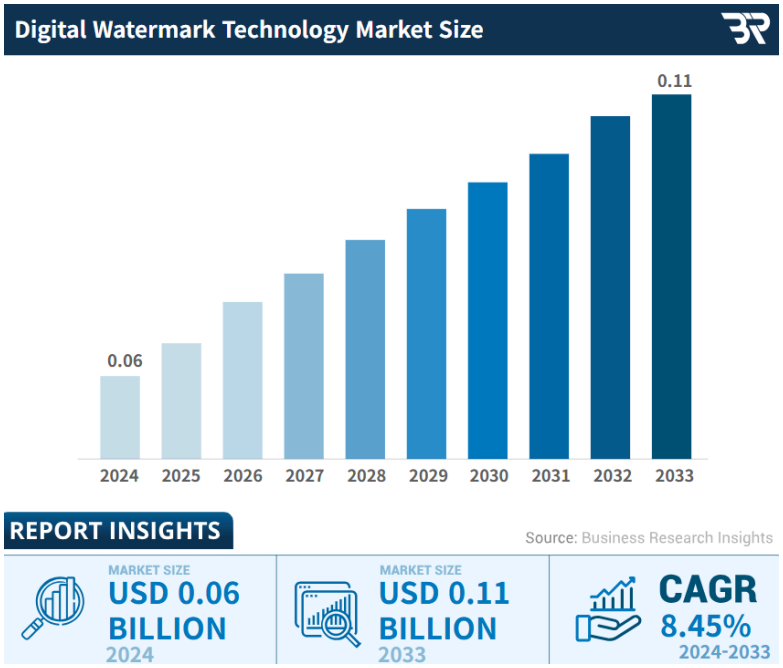
■ 구성 목적: **AI** 모델 결과물**,** 이미지**,** 로그 등 정적 파일을 안전하게 저장하고, 대용량 데이터를 효율적으로 관리하기 위함.

③ 활용방안**(**공공성**)** 및 기대효과

시장 기회**:** 신뢰 붕괴가 야기한 거대 리스크 시장

생성형 AI는 '디지털 신뢰' 자체를 붕괴시키며, 모든 산업에 걸쳐 신뢰 검증은 더 이상 선택이 아닌 필수로 자리잡음. 과거 '저작권 보호'에 머물렀던 워터마크 시장은 이제 AI로 생성된 허위 정보, 딥페이크 사기, 평판 공격 등을 방어해야 하는 더 크고 시급한 문제에 직면함.

시장조사 기관 비즈니스 리서치 인사이트에 따르면, 글로벌 디지털 워터마크 기술 시장 규모는 2033년 까지 11,000만 달러에 도달, 2024년부터 2033년까지 연평균 복합 성장률(CAGR)이 8.45%에 달할 것으로 예상됨.



하지만 기존 기술들은 아래와 같은 명백한 한계로 인해 이 거대한 리스크 시장의 요구를 충족시키지 못하고 있음.

- 기술적 한계**:** 가시성 워터마크는 AI로 쉽게 제거되며, 기존 비가시성 워터마크는 노이즈 삽입 및 재구성 공격에 취약함
- 법적 한계**:** 변조 사실만 알려줄 뿐, 법적 분쟁의 핵심인 '의도성', '구체적 피해 내용', '손해 규모'를 입증할 결정적 증거를 제공하지 못함

본 서비스는 독보적인 기술력(최고 성능 모델 및 모듈화 설계)과 결정적 증거 생성 능력을 결합하여, 신뢰 검증 시장을 재편하고 선도하고자 함.

1. 솔루션의 핵심 활용 방안 및 파급 효과

-

본 서비스의 기술적 독창성은 사회, 산업, 사법 체계 전반에 걸쳐 다음과 같은 구체적인 활용 방안과 기대효과를 창출함.

- 사회적 활용**:** 가짜뉴스 대응 체계 혁신 및 디지털 신뢰 회복
- 활용 방안: 언론사 및 정부 기관이 공식 콘텐츠 배포 시 무결성 워터마크를 적용, 조작 정보 유표시 즉각적으로 무결성 검증 보고서를 공표
- 기대효과: 허위 정보의 진위 여부를 대중이 직접 판단해야 했던 기존의 혼란을 종식. 가짜뉴스 유포 후 진실 규명까지 수일이 걸리던 사회적 혼란을 단 몇 분 만에 잠재워, 국민의 알 권리를 보호하고 건강한 여론 형성 환경을 조성.

  ⇒ 사회 전체의 디지털 신뢰를 회복시키는 핵심적인 매개물

- 산업적 활용**: IP** 자산 보호 및 리스크 관리 비용 절감
- 활용 방안: 웹툰, 디지털 문구, 엔터테인먼트 등 IP 기업의 콘텐츠에 구매자 정보를 각인하여 불법 유출 시 최초 유포자 특정
- 기대 효과: 불법 웹툰 사이트 하나가 월 400억원의 광고 수익을 올리는 등 막대한 피해가 발생하는 시장에서, 유출 경로를 정확히 추적하여 불법 유통의 근원을 차단함.

  ⇒ 창작자의 저작권 수익 손실을 직접적으로 방지하고, 기업이 IP 자산 보호를 위해 지출하던 사후 대응 비용(법무, 모니터링 등)을 획기적으로 절감

- 사법적 활용**: '**디지털 증거**'**의 새로운 표준 제시
- 활용 방안: 명예훼손, 저작권 침해, 동일성유지 등 법적 분쟁에서 무결성 검증 보고서를 객관적이고 과학적인 증거 자료로 제출
- 기대 효과: 기존 기술의 한계였던 증거의 공백을 채움. 본 서비스는 95% 이상의 변조 위치 특정 정밀도와 100%에 가까운 저작권 정보 복구 정확도를 바탕으로 조작의 의도성과 피해 규모를 명확히 입증하는 새로운 차원의 디지털 증거 능력 제공

  ⇒ 소모적인 진실 공방을 줄이고, 사법 체계가 디지털 범죄에 더욱 효과적으로 대응할 수 있는 기술적 기반 마련

2. 본 프로젝트의 사회 적용 및 확산 로드맵

- **1**단계**:** 사회 핵심 인프라의 신뢰 회복
- 적용 대상: 언론사, 금융기관, 정부/공공기관 등 사회적 신뢰가 필수적인 핵심 인프라 영역
- 적용 방식: 콘텐츠 인증 및 무결성 검증 솔루션(SaaS) 형태로 제공하여, 공식 콘텐츠의 신뢰도를 보증하고 위기 대응을 지원

  ⇒ 가장 시급한 사회 문제를 해결하며 기술의 신뢰성과 효과성을 입증하고, 후속 단계 확산을 위한 기반 마련

-

- **2**단계**:** 창작 생태계 보호 및 **IP** 가치 증대

- 적용 대상: 개인 크리에이터, 웹툰/엔터테인먼트 등 IP(지식재산권) 산업 전반
- 적용 방식: 개인용(B2C)과 기업용(B2B) 솔루션을 함께 제공하여 IP 보호 기술의 대중화를 유도. SNS 자동 탐지 및 주간 리포트 기능으로 디지털 자산 보호를 자동화

  ⇒ 모든 창작자가 자신의 권리를 손쉽게 보호하는 환경 조성 및 대한민국 IP 산업의 글로벌 경쟁력 강화 기여

- **3**단계**:** 디지털 신뢰 인프라의 표준화

- 적용 대상: 카메라 앱, 클라우드, SNS 등 대규모 사용자를 보유한 디지털 플랫폼
- 적용 방식: 핵심 기능을 API/SDK 형태로 제공하여, 파트너사 서비스에 '무결성 인증' 기능을 표준처럼 내재화

  ⇒ 디지털 생태계 전반에 신뢰를 기본값으로 탑재하는 기술 인프라로 자리매김하여, 모든 사용자가 안전하게 콘텐츠를 생성하고 소비하는 환경 조성

-