사단법인 정보인권연구소 세미나

2024년 2월 26일(월) 오후4시

국가인권위원회 10층 배움터

사단법인 정보인권연구소
Institute for Digital Rights

세미나 개요

ㅇ 제목 : 사단법인 정보인권연구소 세미나
- 주요국가 인공지능 규제 정책의 주요내용과 시사점 -

ㅇ 일시 : 2024년 2월 26일(월) 오후4시 - 6시

ㅇ 장소 : 국가인권위원회 10층 배움터

ㅇ 순서

4:00 ~ 4:10	개회	* 사회: 김기중 (정보인권연구소 이사, 법무법인 동서양재 변호사)
4:10 ~ 5:10	발제	유럽 및 미국의 AI 규제 동향 및 주요 쟁점 | 오병일 (정보인권연구소 연구위원, 진보네트워크센터 대표)
		인공지능에 대한 규율 - 표준, 사양, 인증 | 이은우 (법무법인 지향 변호사)
5:10 ~ 5:40	토론	송경호 (연세대 정치학과 BK21 박사후연구원)
		이현경 (KISDI 지능정보사회정책연구실, 부연구위원)
		박소영 (국회입법조사처 입법조사관, 변호사)
5:40 ~ 6:00	플로어 토론 및 참석자 전체 토론	

【 발제 1 】

유럽 및 미국의 AI 규제 동향 및 주요 쟁점

오병일

(정보인권연구소 연구위원, 진보네트워크센터 대표)

사단법인 정보인권연구소 세미나

유럽 및 미국의 AI 규제 동향 및 주요 쟁점

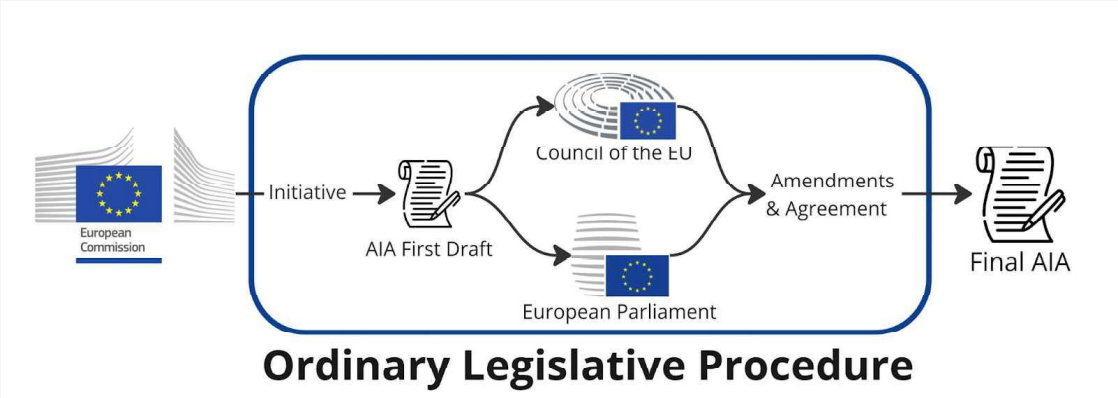
오병일
진보네트워크센터 대표/정보인권연구소 연구위원

유럽연합 인공지능 법안

주요 경과

- 2019.4.8. AI에 대한 고위급 전문가 그룹 : 신뢰할 수 있는 AI 윤리 가이드라인 발표
- 2020.2.19. EC, AI 백서 발표
- 2020.7.17. AI에 대한 고위급 전문가 그룹 : 신뢰할 수 있는 AI에 대한 최종 평가 목록(ALTAI) 발표
- 2021.4.21. EC, AI Act 제안 발표
- 2022.12.6. 유럽연합 이사회(Council of European Union), AI Act에 대한 공통 입장문(일반적 접근) 채택
- 2023.6.14. 유럽의회 AI Act에 대한 수정안(협상안) 채택
- 2023.12.9. EC, 이사회, 유럽의회 3자 협상을 통해 인공지능 법안에 대해 잠정 합의
- 2024.2.2. 이사회에서 만장일치로 통과
- 2024.4.10-11 유럽의회 표결 예정

유럽연합 입법 절차



출처 : https://artificialintelligenceact.eu/context/

인공지능의 정의

다양한 수준의 자율성을 가지고 작동하도록 설계되고, 배치 후 적응성을 나타낼 수 있으며, 명시적 또는 암묵적 목적을 위해 수신된 입력으로부터 물리적 또는 가상 환경에 영향을 미칠 수 있는 예측, 내용, 권장 사항 또는 결정과 같은 출력을 생성하는 방법을 추론하는 기계 기반 시스템

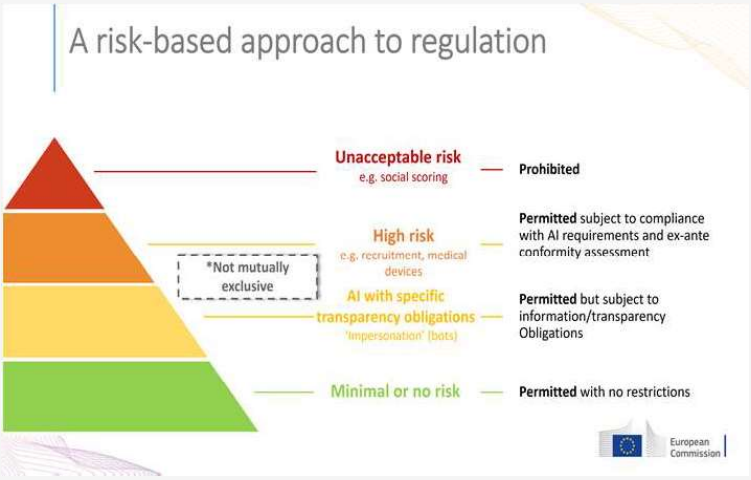
'AI system' means a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments

- 국제적인 조화와 폭넓은 수용을 위해 OECD의 인공지능 정의 차용
  - OECD 정의 : 사람이 정의한 특정 목표 집합에 대해 실제 또는 가상 환경에 영향을 미치는 예측, 권장 사항 또는 결정을 내릴 수 있는 기계 기반 시스템
- 범용 AI 및 생성형 AI 환경 반영

AI Act 적용 범위

- 유럽연합 내에 설립되었는지에 관계없이 유럽연합 내에 AI 시스템 혹은 범용 AI 모델을 출시하거나 서비스를 제공하는 제공자
- 유럽연합 내에 설립되었거나 위치한 AI 시스템 배치자(Deployer)
- 시스템에 의해 생산된 결과물이 유럽연합에서 사용되는 (제3국에 위치하거나 설립지가 있는) AI 시스템의 제공자 및 배포자;
- 적용 제외
  - 오로지 군사 또는 국방 목적으로만 사용되는 시스템
  - 오로지 과학적 연구 및 개발을 위해 사용하는 AI 시스템 및 모델
  - 업무 목적 외로 AI 시스템을 사용하는 자연인(배치자)

AI Act의 위험 기반 접근법



출처 : https://www.spiceworks.com/tech/artificial-intelligence/articles/ai-regulation-best-approach/

금지되는 인공지능 (Title II)

- 잠재의식 기술이나 조작, 기만적인 방법을 사용하여 행동을 왜곡하고 정보에 입각한 의사 결정을 방해하여 심각한 피해를 초래하는 AI 시스템
- 나이, 장애 또는 사회적 또는 경제적 상황으로 인한 취약점을 악용하여 심각한 피해를 야기하는 AI 시스템
- 인종, 정치적 의견, 노동조합 가입 여부, 종교적 신념, 성생활 또는 성적 지향을 유추하는 생체 인식 분류 시스템(법 집행 기관의 합법적인 라벨링 또는 필터링은 제외) - 신설
- 사회적 행동 또는 개인적 특성을 기반으로 개인 또는 그룹을 평가하거나 분류하여 관련 없는 맥락에서 해롭거나 불균형적인 대우를 하거나 행동과 정당하지 않은 대우를 초래하는 AI 시스템
- 법 집행을 위한 공공장소에서의 '실시간' 원격 생체 인식 (피해자 또는 실종자 수색, 안전 위협 방지, 심각한 범죄 용의자 신원 확인 등 특정 필수 목적 제외)
- 프로파일링 또는 성격 특성만을 기반으로 개인의 범죄 위험성을 평가하는 AI 시스템 (범죄 행위와 관련된 객관적이고 검증 가능한 사실에 근거하여 사람의 평가를 지원하는 경우 제외) - 신설
- 인터넷이나 CCTV 영상에서 비표적 스크래핑을 통해 얼굴 인식 데이터베이스를 생성하는 AI 시스템 - 신설
- 직장이나 교육 기관에서 감정을 추론하는 AI 시스템(의료 또는 안전상의 이유 제외) - 신설

금지되는 인공지능 조항에 대한 평가

- 다양한 요건을 충족해야 하기 때문에 금지되는 범위가 넓은 것은 아님.
- 금지되는 인공지능 목록의 업데이트 메커니즘 부재 (고위험 인공지능은 제7조에서 업데이트 절차를 규정하고 있음)
- 생체인식 분류 시스템, 개인 식별을 전제로 한 생체인식 정보에 기반한 분류 시스템으로 제한 : 의회안의 (개인식별과 무관한) '생체인식 기반 데이터' 개념은 미도입
- 스크래핑을 통한 얼굴인식 DB 금지 : 미국의 클리어뷰(ClearView) AI와 같은 관행 금지
- 감정인식 시스템 : (의회안에서 금지했던) 법집행 및 국경관리 목적의 AI는 제외 + 의료, 안전 목적 감정인식 AI 허용
- 사회적 점수 시스템 : 공공기관(AI Act 초안) 뿐만 아니라 민간 시스템도 포함
- 공공장소에서 실시간 원격 생체인식
  - 의회안에서는 예외없이 금지하였으나 합의안에서는 제한적 허용(법집행 기관의 요구 수용)
  - 사전에 법원의 허가가 필요하며 엄격하게 제한된 범죄에 대해서만 시행 , 기본권 영향평가를 완료하고 데이터베이스에 등록해야 함
  - 사후 원격 생체인식은 금지가 아니라 고위험

고위험 AI 시스템

- 부속서 II의 유럽연합 조화 법률 관할의 제품 혹은 안전요소 + 제3자 적정성 평가를 받은 경우
- 부속서 III의 고위험 인공지능 시스템
- 다음과 같은 목록이 합의안에 고위험 인공지능으로 추가됨.
  - 민감하거나 보호되는 속성 정보의 추론에 기반한 생체인식 분류 시스템, 감정인식 시스템
  - 적절한 교육 및 직업 훈련 수준 평가 목적 AI 시스템, 시험 중 학생의 금지된 행위 모니터링 AI 시스템
  - 생명 및 건강보험에서 개인에 대한 위험평가 및 가격책정 시스템
  - 자연인을 감지, 인식 또는 식별할 목적으로 이주, 망명 및 국경 통제 관리의 맥락에서 유럽연합 기관 등이 사용하는 시스템
  - 선거 또는 국민투표의 결과 또는 선거 또는 국민투표에서 투표권을 행사하는 자연인의 투표 행위에 영향을 미치기 위해 사용하려는 AI 시스템
- 고위험 인공지능 예외 규정
  - AI 시스템이 의사 결정 결과에 중대한 영향을 미치지 않는 등 자연인의 건강, 안전 또는 기본권에 중대한 위해를 가할 위험이 없는 경우
- AI 시스템이 자연인 프로파일링을 수행하는 경우 AI 시스템은 항상 고위험으로 간주됨

고위험 AI 시스템의 의무

- 고위험 AI 시스템의 제공자는 △위험 관리, △데이터 세트의 품질 관리를 위한 데이터 평가와 데이터 거버넌스, △기술 문서화 및 기록, △인공지능을 사용하는 자에 대한 정보 제공, △인적 감독, △견고성·정확성·사이버 보안의 요구사항 준수 등 의무
- 고위험 AI 시스템의 제공자는 출시 전에 △품질 관리 시스템을 구축하고 시판 후 모니터링 시스템을 구축하며, △기술 문서 및 로그 기록을 작성하고, △필요한 적합성 평가 절차를 이행하고 CE 인증을 받아야 함.
- EU 고위험 데이터베이스에 등록하고 규제기관에 협력해야 함
  - 고위험이 아니라고 판단하는 AI 시스템 제공자는 해당 시스템을 시장에 출시하거나 서비스에 투입하기 전에 그 평가를 문서화해야 하고, EU 데이터베이스에 등록하며, 당국이 요청할 경우 평가 문서 제출해야 함.

범용 AI에 대한 규율

- AI Act 초안 발표 이후 챗GPT 등 범용 AI 확산 : 범용 AI에 대한 규율 여부 및 방안이 큰 쟁점이 됨.
- 범용 AI(General Purpose AI, GPAI) 모델 : 대규모 자기지도 학습(self-supervision)을 사용하여 대량의 데이터로 학습된 경우를 포함하여 상당한 일반성을 나타내며 모델이 시장에 출시되는 방식에 관계없이 광범위한 고유 작업을 능숙하게 수행할 수 있고 다양한 다운스트림 시스템 또는 애플리케이션에 통합될 수 있는 AI 모델
- 연구, 개발, 프로토타입 제작 활동을 위해 시장에 출시되기 전에 사용되는 AI 모델에는 적용되지 않음.
- GPAI 모델 의무
  - 기술 문서(훈련 및 테스트 절차 및 결과) 유지
  - 해당 모델을 사용하는 AI 시스템 제공업체에 정보 제공
  - 유럽위원회 및 국가 당국과 협력
  - 훈련 콘텐츠의 충분히 상세한 요약본 공개
  - 저작권법 준수 : 유럽연합 디지털 단일시장(DSM) 저작권 지침(Directive 2019/790)에 따라 권리자는 과학적 연구 목적이 아닌 한 텍스트 및 데이터 마이닝을 방지하기 위해 저작물 또는 기타 주제에 대한 권리를 유보할 수 있음. 이 경우 범용 AI 모델 제공자는 해당 저작물에 대해 텍스트 및 데이터 마이닝을 수행하려는 경우 권리자의 승인을 받아야 함.

시스템적 위험이 있는 범용 AI에 대한 규율

- 유럽연합 수준에서 시스템적 위험
  - GPAI 모델의 영향력이 높은 기능(high-impact capabilities)에 고유한 위험
  - 공공건강, 안전, 보안, 기본권, 사회전체에 실제 혹은 합리적으로 예견된 부정적 영향을 국내 시장에 상당히 크게 미치는 경우.
  - 학습에 상당한 양의 계산 능력(부동 소수점 연산으로 측정하여, 총 계산 능력이 10^25를 초과하는 경우)이 포함되는 경우 시스템적 위험이 있는 것으로 분류됨
- 제공자는 이러한 기준을 충족하는 경우 유럽위원회에 통보. 다만, 시스템적 위험을 초래하지 않는다는 근거를 제시할 수 있음.
- 제공자의 의무
  - (시스템적 위험을 식별하고 완화하기 위해 적대적 테스트를 수행하고 문서화하는 것을 포함하여) 표준화된 모델 평가를 수행
  - 심각한 사고를 추적 및 보고
  - 적절한 사이버 보안 보장

AI 거버넌스

- ECI내에 AI 사무국(Office) 설치
- European Artificial Intelligence Board (유럽 인공지능 이사회) 설립
  - 회원국당 한 명의 대표로 구성. EDPS 참관 자격. 이사회 의장은 회원국 대표 중 한명. 유럽 AI 사무국이 이사회 사무국 제공.
  - 임무 : AI 법의 일관되고 효과적인 적용을 위해 EC와 회원국에 조언 및 지원 (58조)
- 자문포럼(58a조)
  - 이사회와 EC에 기술 전문 지식을 조언하고 제공.
  - 업계, 시민사회, 학계 등 이해관계자 대표로 구성. 기본권청, 유럽연합 사이버보안청, 유럽표준화위원회(CEN), 유럽전기기술표준화위원회(CENELEC) 및 유럽전기통신표준협회(ETSI)는 자문 포럼의 영구 회원
- 독립전문가 과학패널(58b조)
  - 인공지능 전문가로 구성. 유럽AI 사무국에 조언 및 지원 : 범용 AI 모델 및 시스템, 시장감시당국업무
- 국가관할당국의 지정 및 단일연락소(59조)
  - 하나의 지정기관(notifying authority) 및 시장감시기관을 국가관할기관으로 지정해야 함.

AI 거버넌스에 대한 평가

- 초안은 유럽연합 차원의 협력과 조율을 위해 각 국가 감독기관과 유럽개인정보보호감독관(EDPS)으로 구성되는 유럽인공지능이사회 설립 : 의장은 EC가 맡고 자문 역할에 한정
- 유럽의회는 법인격을 갖는 독립적 기구로 유럽 인공지능 사무소(European Artificial Intelligence Office)의 신설 제안 : 사무소에 운영이사회, 사무처, 자문포럼을 두도록 하고 있으며, EC에 대한 자문 및 지원 역할과 함께 훨씬 적극적이고 다양한 역할을 수행.
- 합의안의 AI 사무소는 EC 내 기능일 뿐. AI 이사회는 EC에 대한 자문 역할. 의회안보다는 EC의 주도적인 역할을 인정하는 방향으로 타협.

AI 시스템 배치자(사용자)의 의무

- 초안의 사용자(user)를 배치자(deployer)로 변경
- 고위험 인공지능 배치자의 의무 (29조)
  - 인적 감독 할당 및 필요한 지원 보장
  - 입력데이터의 관련성 및 대표성 확인
  - 사용 지침에 따라 운영 모니터링
  - 자동생성로그 보관
  - 자연인과 관련한 결정을 내리는 경우 당사자에게 고지
  - 개인정보 영향평가 및 기본권 영향평가 수행
- 배치자의 기본권 영향평가 신설: 공공기관, 공적 서비스를 제공하는 민간기업, 부속서 III 5조 b,d 운영자(신용평가, 보험평가)
- 고용주의 의무 : 직장에 AI를 배치하려는 조직의 경우, 고위험 AI 시스템을 직장에서 서비스하거나 사용하기 전에 고용주인 배치자는 근로자 대표와 해당 근로자에게 해당 시스템이 적용될 것임을 알려야 함.
- 공공기관 고위험AI 배치자는 제51조 등록 의무 준수 필요

투명성 의무

- 자연인과 직접 상호 작용하도록 의도된 AI 시스템의 경우, 자신이 AI 시스템과 상호 작용하고 있음을 알려야 함
- 워터마킹 : 합성 콘텐츠의 경우 해당 콘텐츠가 인위적으로 생성, 조작되었음을 기계가 읽을 수 있는 형식으로 출력물에 표시
  - 법률에서 범죄를 탐지, 예방, 수사 및 기소를 위한 AI 시스템은 예외
- 감정인식 시스템, 생체인식 분류 시스템의 배포자는 이에 노출된 사람에게 시스템 운영에 대해 알려야 함.
- 딥페이크 : 배치자는 해당 콘텐츠가 인위적으로 생성 또는 조작되었다는 사실을 공개해야 함.
  - 콘텐츠가 명백히 예술적, 창작적, 풍자적, 허구적인 저작물인 경우, 저작물의 전시 또는 향유를 방해하지 않는 적절한 방식으로 구현

영향을 받는 사람의 권리

- 인공지능의 영향을 받는 사람(affected person) 개념 도입
  - 제공자 - 배치자 - 영향을 받는 사람
- 시장 감시 기관에 불만(민원)을 제기할 권리
  - 본 규정의 조항이 침해되었다고 판단할 근거가 있는 자연인 또는 법인은 관련 시장 감시 기관에 불만 사항을 제출할 수 있음.
  - 불만사항은 시장 감시 활동을 수행할 목적으로 고려되어야 하며 이에 따라 시장 감시 당국이 확립한 전용 절차에 따라 처리되어야 함.
- 개인의 의사결정에 대한 설명을 받을 권리.
  - 영향을 받는 모든 사람에게 건강, 안전 및 기본권에 악영향을 미친다고 생각하는 방식으로 법적 영향을 미치거나 유사하게 영향을 미치는 경우 배치자에게 의사 결정 절차에서 AI 시스템의 역할과 결정의 주요 요소에 대해 명확하고 의미 있는 설명을 요청할 권리 부여
  - GDPR이 '오로지 자동화된 처리에만 의존하는 결정'으로 엄격하게 규정하고 있는 것에 비해 폭넓게 권리 인정

벌칙

- 금지된 AI 시스템 의무 미준수 : 최대 3500만 유로 또는 총매출액의 최대 7%
- 고위험 시스템의 요건 미준수 : 최대 1500만 유로 또는 총매출액의 최대 3%
- 부정확하거나 불완전한 정보 제공 : 최대 750만 유로 또는 총매출액의 최대 1%

다음 단계

- 공식 저널 발표 후 20일 후에 발효
- 발효 후 6개월부터 금지되는 인공지능 관행 적용
- 발효 후 12개월부터 범용 AI 모델 의무 적용
- 다른 조항은 발효 후 24개월 후부터 적용
- 제3자 적합성 평가 대상인 규제 대상 제품의 안전 요소인 경우 발효 후 36개월 후부터 적용
- 실천 강령(code of practice)은 발효 후 9개월 이내에 준비되어야 함.

유럽연합 AI Act 시사점

- 세계 최초로 인공지능을 포괄적으로 규율하는 법
  → 인공지능 제품과 서비스의 국제적인 성격을 고려할 때 전 세계 다른 국가에 규범화 효과(브뤼셀 효과)를 가질 것
  → 다만, 실제 적용되는 것은 사실상 2~3년 후이기 때문에 규제가 유예될 우려
- 인권보호 측면에서 합의안은 시민사회의 제안 및 유럽의회안에 비해 후퇴함
  → 프랑스, 독일, 이탈리아 등이 산업계 입장 대변, 각 국의 법집행 당국의 입장 고려
  → 유럽 시민사회는 합의안의 한계에도 불구하고 통과해야 한다는 입장임. 향후 유럽 정치 지형이 더욱 극우화할 것을 우려.
- 산업계는 유럽의 인공지능 산업이 발전하지 못하기 때문에 강력한 규제 체제를 만드는 것이라고 주장
  → 부차적 고려사항은 되겠지만, 규제의 필요성 여부와 효과적인 규제 방식에 초점을 맞춰야 함
  → 인공지능 산업 발전을 위해 인권 침해를 방치하자는 것이 아니라면!

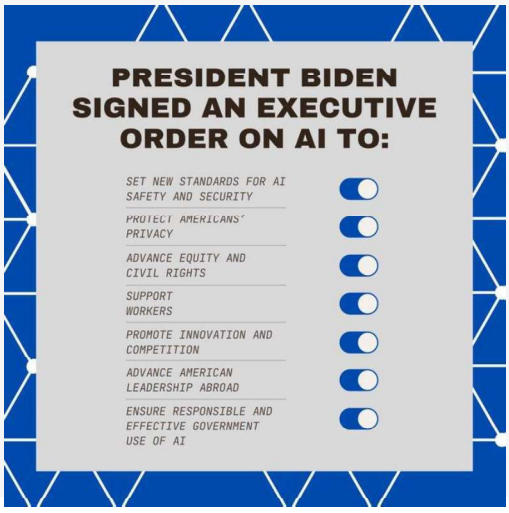
미국 인공지능 행정명령

주요 경과

- 2019.2.11. 트럼프 정부, 인공지능에서 미국의 리더쉽 유지 행정명령 발표 (행정명령 13859)
- 2020.12.3. 트럼프 정부, 연방정부에서 신뢰할 수 있는 인공지능 사용 증진 행정명령 발표(행정명령 13960)
- 2021.1.12. 백악관, 2020년 국가인공지능이니셔티브법에 따라 국가인공지능이니셔티브실 설립
- 2022.10. 백악관 과학기술정책실(OSTP), 인공지능 권리장전을 위한 청사진 발표
- 2023.1.26. 국립표준기술연구소(NIST), AI 위험관리 프레임워크 (AI RMF 1.0) 발표
- 2023.7.21. 바이든 정부, 주요 AI 기업으로부터 AI 위험성을 관리하겠다는 자발적인 약속 받아냄
- 2023.10.30. 바이든 정부, 안전하고 보안이 되며 믿을 수 있는 인공지능의 개발 및 이용에 대한 행정 명령 발표 (행정명령 14110)
- 2023.11.1. 관리예산실(OMB), 인공지능 행정명령에 대한 이행 가이드 발표

인공지능 행정명령 주요 내용

- 목적 : AI의 개발과 사용을 안전하고 책임감 있게 관리하는 것이 가장 시급한 과제이며, 연방정부 차원의 조정된 접근 방식이 필요함.
- 8가지 지도 원칙
  - 안전과 보안
  - 혁신과 경쟁 촉진
  - 근로자 지원
  - 평등과 시민권 보호
  - 소비자 보호
  - 개인정보보호
  - 연방정부의 AI 활용 증진
  - 해외에서 미국의 리더십 강화



인공지능의 안전과 보안

- 국방물자생산법(DPA)에 따라, 강력한 AI 시스템 개발자가 안전 테스트 결과와 기타 중요 정보를 미국 정부와 공유 요구
  - 연방정부 제공을 위한 이중용도 기초모델 개발업체, 대규모 컴퓨팅 클라스터 업체
  - 대규모 AI 모델 훈련을 위해 외국인이 미국의 IaaS 제공업체와 거래할 때, 외국인의 신원과 AI 모델에 대한 정보 제공 요구
- AI 시스템의 안전과 보안, 신뢰성을 보장하기 위한 지침, 표준, 모범사례 개발
  - NIST 표준, 국토안보부 AI 안전 및 보안위원회 설립, CBRN(화학, 생물학, 방사능, 핵) 위협에 대처
- 위험한 생물학적 물질 제조에 AI를 사용할 경우의 위험 관리
- AI 생성 콘텐츠에 대한 인증 및 워터마킹[라벨링] 지침 개발
- 중요 인프라보호와 사이버보안 개선

혁신과 경쟁 촉진

- 인재 유치 : AI에 전문성이 있는 이민자와 비이민자가 미국에서 공부하고 일할 수 있는 기회 확대
- 혁신 촉진 : AI 연구자와 학생들에게 주요 AI 리소스와 데이터에 대한 엑세스를 제공하여 연구 촉진
- 경쟁 촉진 : 소규모 개발자와 중소기업을 지원하여 공정하고 개방적이며 경쟁력있는 AI 생태계 촉진

근로자 지원

- 인공지능은 미국의 일자리와 직장을 변화시키고 있으며, 생산성 향상에 대한 가능성과 함께 직장 내 감시, 편견, 일자리 이동의 위험성 야기
- 이러한 위험을 완화하고 근로자의 단체 교섭 능력을 지원하며, 모두가 접근할 수 있는 인력 교육 및 개발에 투자할 필요
- 일자리 대체, 노동 기준, 직장 내 형평성, 건강 및 안전, 데이터 수집에 관한 원칙과 모범 사례를 개발
- AI의 잠재적 노동시장 영향에 대한 보고서 작성 및 노동 중단 근로자를 위한 연방 정부의 지원 강화 방안 연구

평등과 시민권 보호

- 형사 사법 시스템에서 AI 및 시민권 강화
  - AI 관련 민권 침해 조사 및 기소를 위한 모범 사례에 대한 교육, 기술 지원, 법무부와 연방 민권 사무소 간의 조정을 통해 알고리즘 차별 해결
  - 선고, 가석방 및 보호 관찰, 재판 전 석방 및 구금, 위험 평가, 감시, 범죄 예측 및 예측 치안, 포렌식 분석에 AI를 사용하여 시민권 보호와 법집행 효율성을 향상할 수 있는 모범 사례 개발
- 정부 혜택 및 프로그램과 관련된 시민권 보호.
  - AI 알고리즘이 차별을 악화시키지 않도록 임대인, 연방 혜택 프로그램, 연방 계약업체에 명확한 지침 제공
- 더 넓은 경제에서 AI와 시민권 강화
  - 채용, 주택, 금융에서 AI를 통한 차별 및 편견 방지

소비자 보호

- 사기, 차별, 개인 정보 보호 위협 및 다른 AI 사용 위험으로부터 미국 소비자 보호
- 환자 보호 : 의료분야에서 AI의 책임감 있는 사용과 생명을 구할 수 있는 저렴한 약품 개발을 촉진
- 승객 보호 : 운송 부문에서 AI의 안전하고 책임 있는 개발 및 사용 촉진
- 학생 보호 : AI 기반 교육 도구 등 교육을 혁신할 수 있는 AI의 잠재력을 구체화
- 이용자 보호 : AI가 통신 네트워크와 소비자에게 미치는 영향 검토

개인정보 보호

- 훈련 데이터의 개인정보 보호 등 개인정보 보호 기술(PET)의 개발과 사용을 가속화하기 위한 연방 정부의 지원
- 암호화 도구와 같은 프라이버시 보호 연구 및 기술을 강화
- 연방기관이 상업적으로 이용 가능한 정보를 처리하는 방법을 평가하고 AI 위험을 고려할 수 있도록 개인정보 보호 지침 강화
  - 개인정보 보호 위험을 완화하는 데 개인정보 영향평가가 어떻게 더 효과적일 수 있는지에 대한 의견 요청
- 개인정보 보호 기술의 효과를 평가할 수 있는 지침 개발
- 초당적인 데이터 개인정보 보호 법안을 통과시킬 것을 의회에 촉구 (Fact sheet)

연방정부의 AI 활용 증진

- AI를 통한 정부 서비스의 효율성 확대, 그러나 차별과 안전하지 않은 결정과 같은 위험 존재
  - 연방기관의 AI 사용에 대한 지침 발생
  - 각 기관은 최고인공지능책임자 지정
  - 각 기관 내부에 인공지능 거버넌스 위원회 신설
  - AI 권리장전 청사진 및 NIST 위험관리 프레임워크에 기반한 위험관리 관행 수립 : 공개 의견수렴, 데이터 품질 평가, 서로 다른 영향과 알고리즘 차별 평가 및 완화, AI 사용 통지, 사용하고 있는 AI에 대한 지속적인 모니터링 및 평가, 인간의 고려, AI의 불리한 결정에 대한 구제수단 제공 등
  - 연방 직원의 업무에 생성 AI를 사용하는 방법에 대한 지침 개발
- 보다 효율적인 계약을 통해 특정 AI 제품과 서비스를 효과적으로 획득할 수 있도록 지원
- AI 전문가의 신속한 채용과 직원을 대상으로 한 AI 교육 제공

해외에서 미국의 리더십 강화

- AI 협력을 위한 양자, 다자, 다중 이해관계자 참여 확대
  - AI의 위험을 관리하고 안전을 보장하기 위한 강력한 국제 프레임워크 구축 노력 주도
- 국제 파트너 및 표준 기구와 함께 중요한 AI 표준의 개발 및 구현을 가속화
- 안전하고 책임감 있고 권리를 보장하는 AI의 해외 개발 및 사용을 촉진
  - NIST, AI 위험 관리 프레임워크의 원칙을 통합하는 AI 글로벌 개발 플레이북 발행
- 글로벌 AI 연구 의제를 개발
- 중요 인프라에 대한 국경 간 및 글로벌 AI 위험 해결하기 위해 국제 동맹국 및 파트너와 협력

인공지능 행정명령의 시사점

- 트럼프 정부에서의 행정명령이 AI의 활용과 윤리적 원칙의 표명에 그쳤다면, 바이든 정부의 행정명령은 AI의 부정적 영향에 대한 인식과 AI의 책임감 있는 사용을 위한 대책에 초점
- 글로벌 AI 규범 형성에 미국의 리더쉽 회복의 의지 표명
  - 개인정보보호(GDPR), 디지털 시장 독점 규제(DMA) 등 분야에서는 유럽연합이 리더쉽 발휘 & 유럽 AI Act 타결 예정
  - 11.1 영국 AI 안전 정상회의 직전에 행정명령 발표
- 법률이 아닌 행정명령의 한계
  - 대규모 AI에 대한 보고 의무 : 국방물자생산법(DPA) 활용 논란 - 현재 진정으로 국가안보적 의미가 있는지, 국방물자생산법 취지에 맞는지.
- 미국은 자율 규제를 위해 AI 법안 제정을 주저하고 있는가?
  - 행정명령에서 바이든 행정부는 의회와 협력하여 미국이 책임 있는 혁신을 선도할 수 있도록 초당적 입법을 추진하겠다는 의지 표명
  - 2023.7.21. 7개 AI 기업의 자발적 약속(voluntary commitments)에서 이는 구속력 있는 의무의 개발 및 집행을 위한 첫 단계일 뿐이며, 근본적으로 같은 이슈를 다루는 초당적 입법을 추진하겠다고 밝힘.

종합 의견

종합 의견

- 인공지능에 대한 포괄적 규율 vs 부문별 규율
  - 유럽 AI Act 는 AI의 위험성을 통제하기 위한 포괄적 법안이고, 미국의 행정명령은 AI 환경에 대응한 부문별 정책 패키지임. (위상이 다름)
  - AI의 위험성 규율과 관련해서는 포괄적인 프레임워크와 특수성에 대한 고려가 모두 필요함. 예를 들어, 유럽 AI Act도 범용 AI, AI의 위험성, 범죄수사 목적 등 특수성을 고려한 규정을 두고 있음. 반면, 미국 NIST의 위험관리 프레임워크 역시 포괄적 프레임워크임.
- 법적 규제 vs 자율규제
  - 2023년을 기점으로 전 세계적으로 AI의 위험성 규제 필요성에 대한 공감대 형성
  - 미국 역시 AI를 규율할 수 있는 입법 추진
- AI 산업 육성 vs AI 위험성 규제
  - AI의 안전성에 대한 신뢰 없이는 AI 산업 발전도 불가능
  - 규제의 목적은 기업에게 부담을 주고자 하는 것이 아니라, 공익과 인권의 보호임
  - AI 규제에 대한 산업 위축론은 논점 왜곡 → AI의 위험성을 식별하고 이에 대한 적절한 규제 체제가 무엇인지에 대한 논의에 초점을 두어야 함

종합 의견

- 인공지능에 대한 국제적인 규율 필요
  - AI 제품 및 서비스의 보급, 그리고 AI가 야기하는 위험성은 본질적으로 국제적 성격
  - 실제로 AI 윤리 및 AI 위험 관리 관행의 내용은 상당히 유사 : 미국, 유럽 모두 위험 기반 접근방식 채택
  - 유럽 및 미국의 규제 사례는 한국에 참조가 될 수 있음
- 한국의 AI 규제
  - 채용 AI 등 공공기관은 이미 민간의 AI 시스템을 도입하고 있지만, 위험성 및 성능에 대한 평가, 책무성과 투명성을 위한 자료 보관, AI 시스템의 공공조달 정책이 존재하는가
  - 국회에 계류되어 있는 인공지능 기본법안의 내용에 대해 충분한 사회적 논의가 이루어졌는가 : 금지되는/고위험 인공지능을 어떻게 규정할 것인지, 고위험 인공지능 제공자/사용자의 책무는 무엇인지, 인공지능 거버넌스는 어떠해야 하는지 등에 대한 사회적 토론을 한 적이 있는가
  - 인공지능 정책 수립에 있어서 'AI의 영향을 받는 사람 ' 의 목소리는 제대로 고려되고 있는가

감사합니다.

【 발제 2 】

인공지능에 대한 규율
- 표준, 사양, 인증

이은우

(법무법인 지향 변호사)

인공지능에 대한 규율 - 표준, 사양, 인증

'AI 영상 검색 및 대상물 이동경로 추적 솔루션' AI 신뢰성 인증 검토

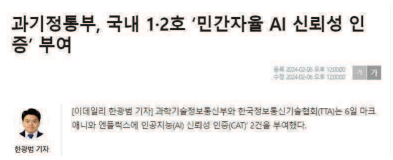
이은우(법무법인 지향)

'AI 시스템'에 대한 규율 추진 현황은

미국, 유럽연합 등 매우 적극적인 움직임

미국	2023. 10. 30. (미국) 바이든 대통령 **"안전하고 신뢰할 수 있는 인공 지능에 대한 행정 명령"**
	2023. 11. 1. (미국) 관리예산처(OMB) Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence
	2023. AI 위험관리 프레임워크(NIST)
	2022. 10. 4. 인공지능 권리장전 청사진(백악관 과학기술정책실)
유럽연합	2023. AI법에 상응하는 AI 표준에 대한 조사(Analysis of the preliminary AI standardisation work plan in support of the AI Act)
	2023. 12. 9. EU 의회, 이사회 AI법 임시합의
	2023. 1. 21. EU 인공지능법 최종 초안

우리나라의 민간 AI 신뢰성 인증 추진 경과


과기정통부, 국내 1·2호 '민간자율 AI 신뢰성 인증' 부여

[이데일리 한광범 기자] 과학기술정보통신부와 한국정보통신기술협회(TTA)는 6일 마크애니와 엔플럭스에 인공지능(AI) 신뢰성 인증(CAT) 2건을 부여했다.

한광범 기자


과기정통부, 초거대AI 신뢰성·성능 평가한다

| 11일 열린 'AI 윤리, 신뢰성 강화 현장간담회'서 밝혀...사실 정확성 등 4개 부문 평가

방은주 기자


과기정통부, AI 신뢰성 검·인증 추진..."인증체계 국제표준화"

AI 영상 검색 및 대상물 이동경로 추적 솔루션

제1호 인공지능 신뢰성 인증

보도자료

보도시점 2024. 2. 6.(목) 12:00 (2024. 2. 7.(수) 조간) 배포 2023. 2. 6.(목) 09:00

국내 1·2호 '민간자율 인공지능 신뢰성 인증' 부여

- 국내 최초 AI 신뢰성 단체표준('23.12.) 기반 인증사례 -
- AI 기술 혁신과 AI 신뢰성 확보의 균형 발전을 위한 인증제도 본격 시동 -

과학기술정보통신부(장관 이종호, 이하 '과기정통부')와 **한국정보통신기술협회**(회장 손승현, 이하 'TTA')는 2월 6일(화) **주식회사 마크애니**(대표이사 최고, 이하 '마크애니'), **주식회사 엔플럭스**(대표이사 이민, 이하 '엔플럭스')에 '**인공지능**(이하 'AI') **신뢰성 인증**(CAT: Certification of Artificial Intelligence)' **2건**을 **부여**하였다고 밝혔다.

'**AI 신뢰성 인증 제도**'는 **AI 신뢰성 단체표준**을 **기반**으로 하며, **AI 신뢰성**을 **자발적으로 확보**하려는 **민간 AI 사업자**를 **대상**으로 **진행**된다. **민간 인증 전문기관**인 **TTA**가 **AI** 기술을 활용한 **제품·서비스의 위험요인**을 **분석**하고, 위협에 기반하여 **신뢰성 확보**를 위한 **사업자**의 **요구사항 준수 여부**를 평가한다.

< 'AI 신뢰성 인증제도' 개요 >

□ (대상) ① 자발적으로 신뢰성을 확보하려는 일반영역 AI 사업자·개발자
② 과기정통부 AI 지원사업 중 고위험영역 AI에 해당하는 사업
□ (범위) 대상(데이터·모델·시스템 등)에 따라 개발안내서의 15개 요구사항 中 필수 요구사항 선별
□ (절차) 개발안내서 적용법 교육 이후 요구사항 선별하여 TTA에서 시험 실시, 인증서 발급

①사전교육 (TTA-공인final 신청자) ⇨ ②검증항목 선별 ⇨ ③개발안내서 적용 ⇨ ④검·인증

※ 과기정통부와 TTA는 '21년부터 분야별 '신뢰할 수 있는 AI 개발안내서'(이하 '개발안내서') 개발·보급을 통해 AI 신뢰성 확보를 위한 기술 요구사항을 정립하고, 국내 기업이 AI 신뢰성을 확보해 나갈 수 있도록 지원해왔다. **지난해 12월에는 개발안내서 내용을 바탕으로, AI 신뢰성 관련 국제표준인 ISO/IEC TR 24028(신뢰성 개요), ISO/IEC 23894(위험관리), ISO/IEC 22989(용어)와의 국제 호환성을 확보한 국내 최초의 AI 신뢰성 정보통신단체표준인 'AI 시스템 신뢰성 제고를 위한 요구사항'**(이하 'AI 신뢰성 단체표준')**을 제정**하였다.

AI 모델 오류를 위험요소로 식별하고, **AI 모델 편향 제거, AI 시스템 신뢰성 테스트체적 수립,** AI 신뢰성 확보를 위한 **기업의 거버넌스 구성** 등을 검증하였다.

* 과기정통부 'AI 융합 국민안전 확보 및 신속대응 지원사업('23)

스마트 관제 전문기업인 **엔플럭스**의 **'AI 융합 지뢰탐지 모듈 v1.0"**은 **지표투과레이더**(GPR) **이미지**를 판독해 **지뢰 여부, 지뢰 종류 판단** 등 **고수준의 분석기능**을 **제공**하는 **시스템**으로, **지뢰탐지 정확성 오류**와 **지뢰탐지 결과**의 **설명가능성 부재** 등을 **위험요소**로 도출하고, **데이터 구축 방법**의 **객결성**과 **AI 모델**의 **판단결과**에 대한 **설명가능성 확보** 등을 **중점적**으로 **검증**하였다.

* 과기정통부 'AI 융합 프로젝트(AI+X)'의 'AI 융합 지뢰탐지시스템 개발·실증 과제('23)

민간 인증 전문기관인 **TTA**는 **AI 신뢰성 검증**을 **실시**하는 과정에서 **신뢰성 확보를 위한 보완 필요사항**에 대해 **사업자 대상 컨설팅**을 **수행**하며 **AI 제품·서비스**의 **신뢰성이 개선**될 수 있도록 **지원**하였다.

이번 인증은 **지난 10월 민간자율 AI 신뢰성 인증제도**를 **도입**한 이후 **첫 사례**로, **국내 AI 제품·서비스**의 **신뢰성**을 **확보**하기 위한 **민간 자율체계**를 **확립**했다는 점에서 **의미**가 있다.

과기정통부는 **전세계적**으로 **AI 신뢰·안전성 확보**를 위한 **노력이 치열**한 **상황**에서, **전문성 있는 기관**을 통한 **민간자율 AI 신뢰성 인증제도**는 **AI 혁신**과 **안전하고 신뢰할 수 있는 AI 활용**을 **동시에 촉진**할 수 있는 **좋은 모델**이 될 것으로 **기대**하며, **빠른속도로 발전**하는 **생성형 AI 기술·지장 변화**를 **반영**하여 **AI 신뢰성 인증제도**를 **고도화**하고, **AI 신뢰성 인증 모범사례**를 **지속 확대**하여 **국내 AI 산업**의 **AI 신뢰·안전성 기반**을 **강화**해 나갈 계획이다.

담당 부서	과학기술정보통신부	책임자	과 장	남철기 (044-202-6270)
	인공지능기반정책과	담당자	사무관	이재홍 (044-202-6282)
정책 기관	한국정보통신기술협회	책임자	센터장	신종호 (010-3439-5538)
	시산회산센터	담당자	팀 장	배종석 (010-5110-2327)

대한민국 정책브리핑 OPEN

< 연계시스템 활용 시나리오(예시) >

< AI융합 국민안전 실증 전체 흐름도 >

AI 영상 검색 및 대상물 이동경로 추적 솔루션
- 최초의 AI 신뢰성 인증(한국정보통신기술협회 단체표준)

"실시간 원격 생체인식 식별 AI 시스템은 논란의 대상..."

- 실시간 원격 생체인식 식별 인공지능 시스템(Real-time and remote biometric identification systems)
  - 오남용의 가능성
  - 오남용시 기본권 침해 우려가 매우 높은 인공지능 시스템

https://reclaimyourface.eu/

인공지능 행위자, 영향을 받는 자, 지역사회 등에게 공개되고, 참여가 보장되었는가?

자율적인 인증으로 신뢰할 수 있는 인공지능 시스템을 규율하려는 자율규제는 믿을 수 있는가?

이와 같은 방식과 결과를 가져오는 'AI 신뢰성 인증'은 신뢰할 수 있는가?

이를 통해 AI 시스템 규율에 대해 어떤 교훈과 시사점을 얻을 수 있는가?

왜 실시간 생체인증 AI 시스템이 1, 2호 신뢰성 인증을 받았는가?

우리는 지금 "신뢰할 수 있는 인공지능"으로 향하고 있는가?

- AI 시스템 신뢰성 인증
- 실시간 생체 인식
- AI 시스템 규율
- 자율규제
- 표준

AI 행위자로 적극 참여해야

AI 시스템의 특징과 규율의 필요성

인공지능

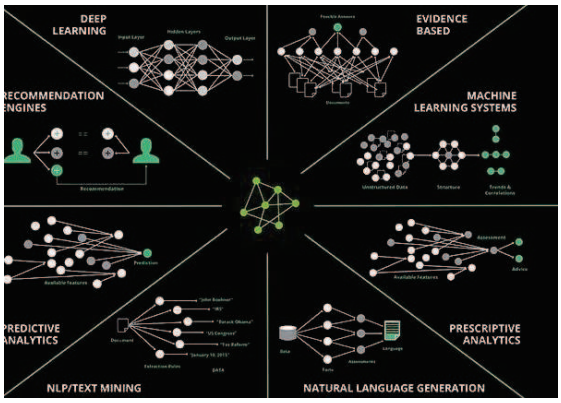
- AI는 경제 발전의 핵심 동인이 될 수 있는 잠재력을 지닌 전략적으로 중요한 영역. AI는 또한 광범위한 사회적 영향을 미칠 수 있음.

- 이에 각국은 기술 및 산업 역량을 강화하고 민간 및 공공 부문 모두 경제 전반에 걸쳐 AI 활용을 강화, AI가 가져올 사회경제적 변화에 대비, 적절한 윤리적, 법적 틀을 보장하는 것을 목표로 함.

인공지능의 특징과 규율 필요성 :

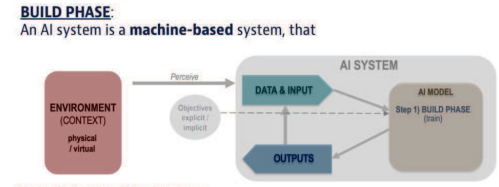
- "기존 소프트웨어에 비해 AI는 다양한 위험을 안고 있음. AI 시스템은 시간이 지남에 따라 때로는 심각하고 예기치 않게 변경되어 이해하기 어려울 수 있는 방식으로 시스템에 영향을 미칠 수 있는 데이터에 대해 훈련. 이러한 시스템은 본질적으로 "사회 기술적", 즉, 사회 역학과 인간 행동의 영향을 받음. AI 위험은 이러한 기술적, 사회적 요인의 복잡한 상호 작용에서 나타날 수 있으며, 온라인 챗봇 경험부터 취업 및 대출 신청 결과에 이르기까지 다양한 상황에서 사람들의 삶에 영향을 미칠 수 있음."

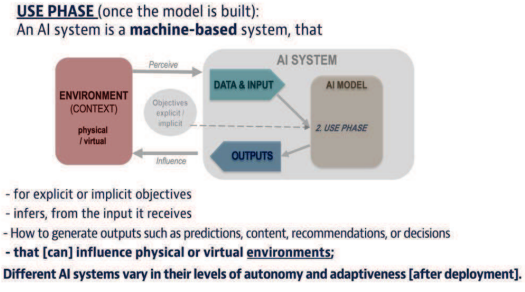
(AI Watch: 인공지능 표준화 환경 업데이트, 유럽 AI 규정의 맥락에서 IEEE 표준 분석 EUR 31343 EN)



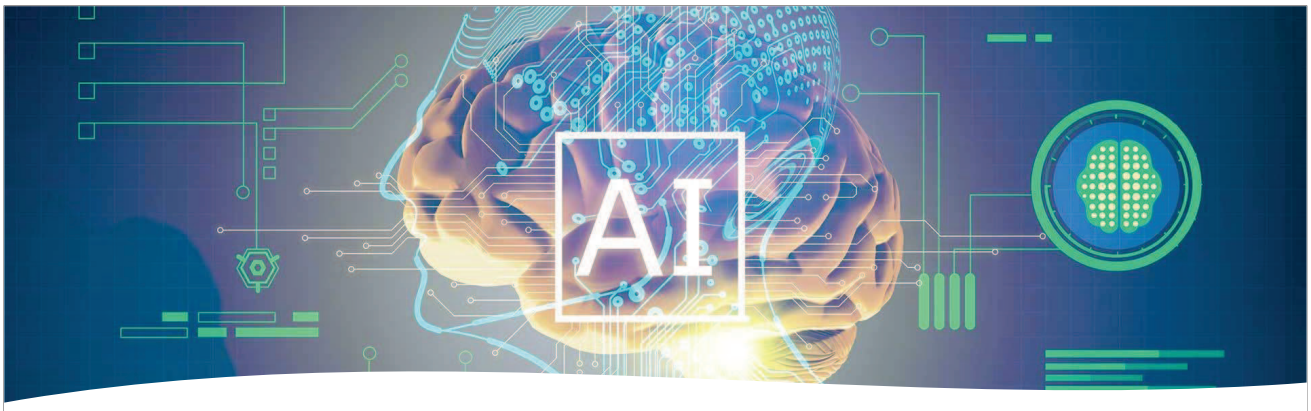
AI 시스템은 폭 넓은 영향을 미침

- 환경과의 상호 작용을 고려해야 함
  - 지역사회에 미치는 영향
- 의사결정에 영향을 미치는 장기적 효과
- 자율성과 적응성에 미치는 장기적 효과
  - 실시간 추적되는 CCTV가 자율성에 미치는 영향





(OECD의 AI 시스템 정의)



AI에 대한 규율

- 종합적인 규율 VS 부분화된 규율
  - 유럽연합
    - AI 법
    - 그 외의 부분적인 법률(DSA, DMA, 표준, 안전 등)
  - 미국
    - 대통령 행정명령(연방정부, 공공부분, 국방, 국가안보 중요분야)
    - 라벨링 의무법 등 개별법 추진
- 법적 구속력이 있는 규율 VS 자율 규제
- 표준, 인증, 사양

(미국)인공 지능의 안전하고 신뢰할 수 있는 개발 및 사용에 관한 행정 명령과 OMB 지침

- 두 문서 모두 엄격한 책임을 요구
  - 연방 기관에 알고리즘 차별로부터 보호하기 위해 민권 보호를 시행하도록 지시.
  - 차세대 AI 모델을 개발하는 회사는 특정 안전, 평가 및 보고 절차를 충족하는지 지속적으로 확인하기 위해 연방 정부에 보고해야 .
- 연방 기관이 기술을 사용하기 위해 준수해야 하는 최소한의 안전 및 권리 보호 기준.
  - 안전 및 권리에 영향을 미치는 AI에 대한 명확한 정의를 제공하고 안전 또는 권리에 영향을 미치는 것으로 추정 되는 특정 시스템 목록을 포함.
  - 채용 알고리즘, 범죄 위험 평가, 의료 AI 장치 등 시스템의 잠재적인 피해로부터 보호하기 위해 영향을 기반으로 가드레일을 만드는 데 중점(OMB 메모 초안)
- 연방 정부를 책임 있는 AI의 모델로 설정
  - 연방 정부가 자체 AI 활용을 규제
  - 법률 제정 이전에 연방 정부는 시장을 형성하고 민간 산업과 잠재적인 미래 입법을 위한 경로를 모델링
- AI
  - AI 권리 장전에 대한 청사진 및 NIST의 2023년 1월 AI 위험 관리 프레임워크 등을 계승. 향후 더 구체화될 것임.

(EU) 인공지능법

- 고위험 AI 시스템 준수사항
  - 요구사항 준수 / 위험 관리 시스템 / 데이터 및 데이터
  - 기술 문서와 기록보관 / 배포자에 대한 정보의 투명성 및 제공
  - 인간의 감독 / 정확성, 견고성 및 사이버 보안
  - 품질경영시스템 / 문서 보관 / 자동 생성 로그
  - 수입업자, 대리점, 유통자, 배포자 의무
  - 기본권 영향 평가
  - 적합성 평가 / 인증 / 조화 표준 / 공통 사양 / 특정 요건에 대한 적합성 추정
- 투명성 의무

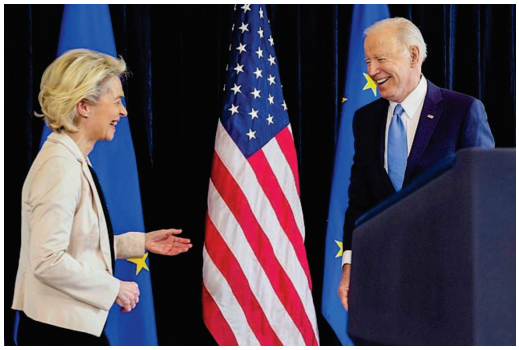
EU-US 공동 로드맵 (신뢰할 수 있는 인공지능에 대한 평가와 측정 도구 및 위험 관리에 대한 공동 로드맵, 2022. 12.)

- AI 표준

"EU와 미국은 기술적으로 건전하고 성능 기반 표준의 개방적이고 투명한 개발을 촉진하여 국제 표준화 노력을 주도하는 것을 목표로 합니다. 시장 경쟁을 위한 일관된 규칙을 확립하고 무역 장벽을 예방하며 혁신을 촉진하려면 국제 AI 표준에 대한 글로벌 리더십과 협력이 필수적입니다. EU와 미국은 국제 표준 개발에 적극적으로 참여하고 WTO 원칙을 준수하며 향후 개발을 위한 격차를 식별함으로써 리더십을 제공하는 것을 목표로 합니다. 이는 이해관계자를 참여시키고, AI 신뢰성, 편견 및 위험 관리를 우선시하며, 프로세스에 중소기업을 포함시킬 것입니다."

신뢰할 수 있는 AI 및 위험 관리를 위한 도구

"EU와 미국은 환경 영향을 포함하여 AI 신뢰성 및 위험 관리를 측정하기 위한 측정 기준 및 방법론의 공유 저장소를 만들기 위해 협력할 것입니다. 다양한 이해관계자의 기존 도구와 표준을 분석하여 공통점, 격차 및 개선 영역을 찾아낼 것입니다. 이러한 연구 결과는 AI 표준 개발에 대한 정보를 제공하고 해당 표준에 부합하는 신뢰할 수 있는 AI 도구의 배포를 촉진할 것입니다."



EU AI법과 표준

- 인공지능법은 "새로운 입법 체계"(NLF)의 원칙 수용
  - 입법자는 필수 요구 사항과 보호 목표를 공식화하고, 유럽 표준 조직에 이를 표준 및 사양의 형태로 기술적으로 지정하도록 요청
  - 고위험 인공지능 시스템이나, 범용 목적 인공지능 시스템이 Regulation (EU) 1025/2012에 따라서 유럽연합 관보에 게재된 '조율된 규격'(harmonized standards)을 충족하는 경우 조율된 규격이 포괄하는 부분의 인공지능법 요구사항을 준수한 것으로 추정해 줌
    - **공인된 유럽 표준 기구(CEN, CENELEC 또는 ETSI)에서 개발한 유럽 표준**
  - 유럽연합 집행위원회는 Regulation EU (No)1025/2012에 따라서 인공지능법의 요구사항을 포괄하는 표준을 제정할 것을 요구하는 요청서를 발행할 수 있음.
  - 특히 고위험 AI 애플리케이션 영역에서는 표준과 사양이 핵심.
  - AI 시스템이 시장에 출시되기 전에 투명성, 정확성, 설명 가능성 또는 품질 측면에서 충족해야 하는 안전 요구 사항을 정의
  - 표준과 사양은 편견, 차별, 조작으로부터 보호하는 데 결정적인 역할을 할 수 있음

EU AI법과 표준

- 법률과 사양(Specifications), 표준(Standards)의 역할
  - 법률 조항에는 기술 수준에서 요구 사항을 충족하는 방법이 명시되어 있지 않음.
  - 대신, **New Legislative Framework[에 따른 유럽 규정의 경우와 마찬가지로 공익을 보호하기 위한 필수적이고 높은 수준의 요구 사항을 정의.**
  - **제품이 이러한 요구 사항을 준수하는 데 필요한 유럽 조화 표준 생성,**
- 이러한 맥락에서 AI법은 **유럽 표준화 기구(ESO)가 작성한 일련의 기술 사양**에 의해 뒷받침될 것
- **사양은 자발적인 성격을 갖고 있지만 법적 요구 사항을 준수한다는 추정을 제공.**
  - **규모와 리소스에 관계없이 모든 AI 제공업체에게 공평한 경쟁의 장을 보장하는 데 근본적인 역할을 할 뿐만 아니라 적합성 평가 절차를 단순화 함.**
- 국제 표준과 기존 표준의 활용
  - AI법을 뒷받침하는 AI 표준과 기술 사양을 정의하는 과정에서 **ESO는 특히 CEN과 ISO 간의 비엔나 협약 또는 CENELEC과 IEC 간의 프랑크푸르트 협약과 같은 국제 표준화 기관과의 협력 협약을 통해 기존 표준 및 기술 사양을 활용**할 수 있음
  - 기존 국제 작업의 채택은 작업 중복을 피하고 다가오는 AI 규정에 필요한 광범위한 표준을 개발하는 데 필요한 시간을 크게 줄이는 가장 효율적인 방법
- **2023. 유럽 위원회는 이미 유럽 표준화 기구에 필요한 표준을 개발하도록 공식 명령을 제공하는 표준화 요청을 채택하는 프로세스를 시작**
- **AI법을 지원하기 위해 채택될 일련의 기술 사양과 기존의 표준 검토 및** AI 법과의 간극 확인
- **윤리적으로 조정된 자율 및 지능형 시스템을 위한 IEEE 7000 시리즈의 표준뿐만 아니라 자율 및 지능형 시스템을 위한 IEEE 윤리 인증 프로그램에서 선택된 인증 기준 제품군, 기존 ISO/IEC 작업과의 비교를 제공하고, 유럽 표준화 작업 내에서 잠재적인 통합을 촉진하기 위해 유럽 요구에 적용해야 할 수 있는 영역을 식별**

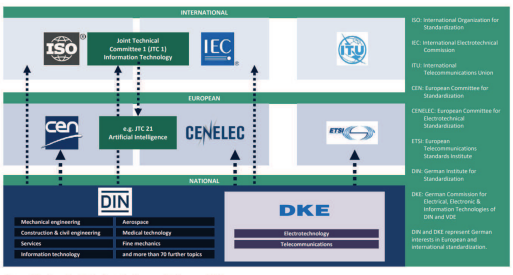


Figure 11: Levels of standardization work (Source: DIN)

표준, 인증, 시장감시 새로운 입법체계(New Legislation Framework)

적합성 평가

- **제조업체는 적용 가능한 모든 요구 사항을 충족하는 경우에만 EU 시장에 제품을 출시할 수 있습니다.**
- 제품이 시장에 출시되기 전에 적합성 평가를 거칩니다.
- 모든 법적 요구 사항을 충족한다는 것을 입증해야 합니다.
- 여기에는 테스트, 검사 및 인증이 포함됩니다.
- 해당 제품 법규에는 각 제품에 대한 절차가 명시되어 있습니다.

적합성평가기관의 인증

- **인증은 유럽 적합성 평가 시스템의 마지막 공공 통제 수준입니다. 적합성 평가 기관(예: 실험실, 검사 또는 인증 기관)이 자신의 임무를 수행할 수 있는 기술적 역량을 갖추어야 합니다.**
- 인증은 공공 부문 활동이자 비영리 활동입니다.
- 국가 인증 기관 간에 경쟁이 없습니다.
- 이해관계자가 대표되어야 합니다

제품에 대한 시장 감시

- **시장 감시는 EU 시장의 비식품 제품이 유럽 소비자와 근로자를 위험에 빠뜨리지 않도록 보장합니다.**
- **이는 또한 환경, 보안, 무역 공정성과 같은 기타 공공 이익의 보호를 보장합니다.**
- **여기에는 규정을 준수하지 않는 제품의 유통을 중단하거나 규정을 준수하도록 하기 위한 제품 회수, 회수 및 제재 적용과 같은 조치가 포함됩니다.**

유럽연합의 조율된 표준과 기술사양

'표준'은 공인된 표준화 기관이 채택한 기술 사양

'기술 사양'은 제품, 프로세스, 서비스 또는 시스템이 충족해야 하는 기술 요구 사항을 규정한 문서

- 품질, 성능, 상호 운용성, 환경 보호, 건강, 안전 또는 치수 수준을 포함하여 제품 판매 이름, 용어, 기호, 테스트 및 테스트 방법, 포장, 표시 또는 라벨링 및 적합성 평가 절차와 관련하여 제품에 요구되는 특성;
- TFEU 제38(1)조에 정의된 농산물, 인간 및 동물 소비용 제품, 의약품과 관련하여 사용되는 생산 방법 및 공정 뿐만 아니라 제품 특성에 영향을 미치는 경우 다른 제품과 관련된 생산 방법 및 공정;
- 품질, 성능, 상호 운용성, 환경 보호, 건강 또는 안전 수준을 포함하여 서비스에 요구되는 특성과 수신자에게 제공되는 정보와 관련하여 공급자에게 적용되는 요구 사항을 포함;

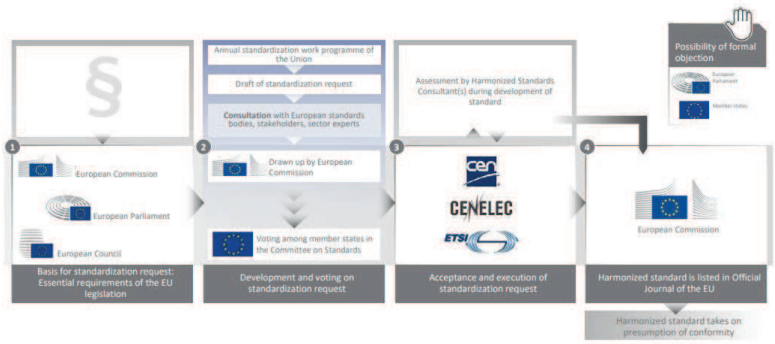
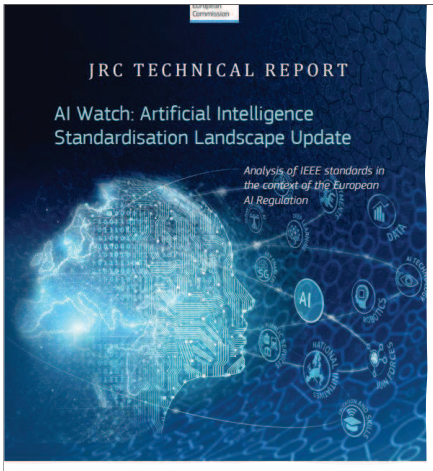
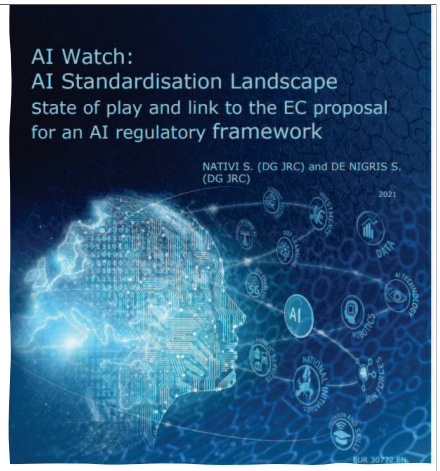


Figure 7: Process of developing harmonized European Standards (Source: DIN)

AI법과 조율된 표준의 관계

- 표준화 요청(Standardization Request)
- 표준 초안
- 표결
- 표준 채택(또는 거부)



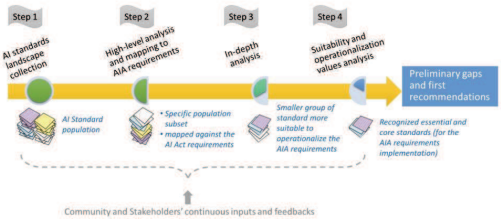


AI법에 맞는 인공지능 표준 작성 추진 중

- **AI 표준화 환경: AI 규제 프레임워크에 대한 실행 상태 및 EC 제안에 대한 링크(2021)**
- **AI Watch: 인공지능 표준화 현황 업데이트(2023)**
- 유럽 표준기관의 주무 기관인 독일 표준기관 : 독일 인공지능 표준화 로드맵(2023)

AI법에 대한 표준화 추진 과정

- 유럽 AI 표준화 로드맵 작성
- **첫 번째 단계는 국제 수준에서 기존 환경을 조사하는 것**
- JRC는 2021년에 **첫 번째 AI 표준화 환경 분석을 제시 -** StandICT.eu 프로젝트에서 생성된 AI 표준화 설문조사와 같은 관련 작업을 기반으로 한 분석은 **주요 국제 및 유럽 표준 개발의 약 140개 표준 및 표준화 결과물**(예: 기술 보고서, 기술 사양 및 인증 기준)을 다룸
- **조직(SDO): ISO/IEC, ETSI, IEEE 및 ITU-T. 주로 당시 최종 또는 초안 형태로 제공되었던 AI 법의 맥락과 관련된 이러한 표준의 하위 집합을 보다 자세히 검토하여 특정 AI 법을 운영할 수 있는 가능성이 있는 유망한 표준의 짧은 목록**
- 요구 사항과 잠재적인 표준화 격차에 대한 예비 목록
- 유럽위원회는 ESO에 표준화 요청
- AI법 제안의 요구 사항을 뒷받침하는 주요 기술 영역을 포괄하고 향후 조화 표준을 위한 기술 기반을 준비하는 역할을 하는 유럽 표준을 요청
- 업데이트 : 협상 결과를 반영
- 표준화 요청을 예상하여 **ESO는 예비 로드맵 정의 활동에 참여했으며 식별된 표준화 요구 사항 중 일부를 해결하는 새로운 사양 개발을 탐색하기 위해 새로운 임시 그룹**을 구성
- **AI 신뢰성 또는 AI 시스템 위험 관리의 기술적 측면 등.**
- 유럽연합 집행위원회는 법적 요구 사항의 관점에서 표준의 기술적 분석을 포함하여 이 프로세스를 지원
- **인공 지능에 관한 JTC1 SC42 - 현재 진행 중인 상당한 양의 ISO/IEC 표준화 활동에도 불구하고 AI 법의 요구 사항을 완전히 포괄하기 위해 해결해야 할 격차가 상당함을 확인.**



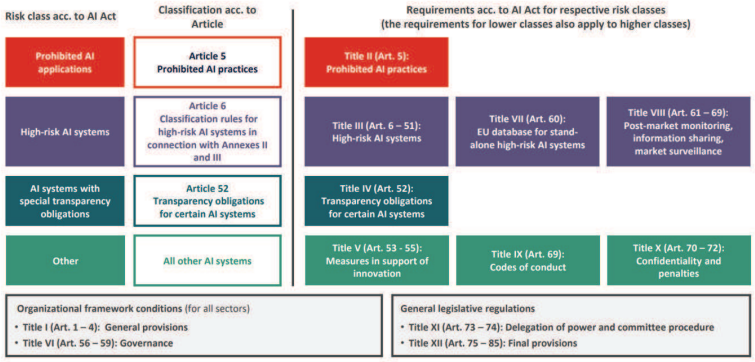


Figure 9: Overview of the content of the planned AI Act (Source: Martin Haimerl)

AI 법과 상응하는 표준화 부분

1. 리스크 관리 2. 데이터 및 데이터 3. 로깅 기능을 통한 기록 보관 4. 사용자를 위한 투명성 및 정보 5. 인간의 감독 6. AI 시스템의 정확도 사양 7. AI 시스템의 견고성 사양 8. AI 시스템을 위한 사이버보안 사양 9. 시판 후 모니터링 프로세스를 포함한 AI 시스템 제공업체를 위한 품질 관리 시스템 10. AI 시스템 적합성 평가

적합성 인증의 절차

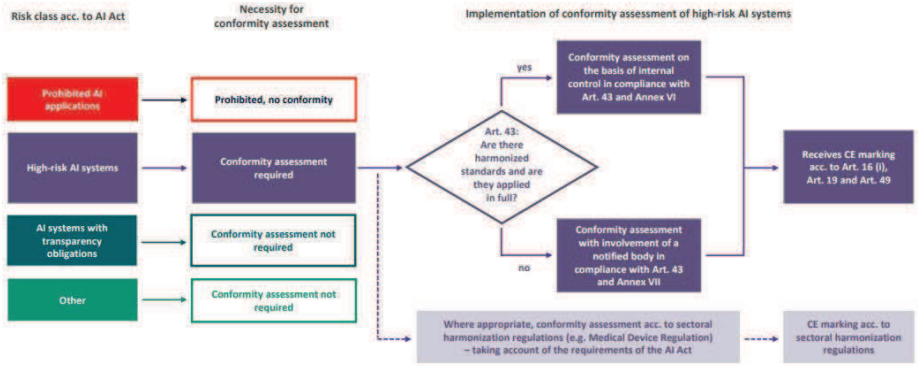


Figure 10: Variations of conformity assessment according to the AI Act (Source: Martin Haimerl)

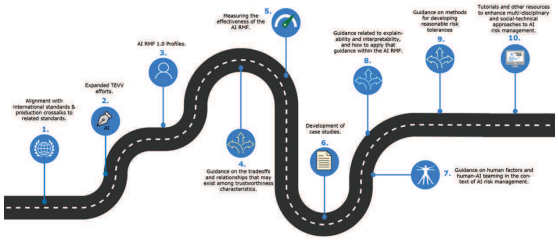
Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment

신뢰할 수 있는 인공지능 자체 평가 리스트

- 인간 대리인 및 감독: 기본권, 인간 대리인 및 인간 감독.
- 기술적 견고성 및 안전성: 공격 및 보안에 대한 복원력, 대체 계획 및 일반적인 안전성, 정확성, 신뢰성 및 재현성.
- 개인 정보 보호 및 데이터 : 개인 정보 보호, 데이터 품질 및 무결성, 데이터 액세스에 대한 존중.
- 투명성: 추적성, 설명성, 의사소통.
- 다양성, 비차별, 공정성: 불공정한 편견의 방지, 접근성 및 보편적 디자인.
- 사회 및 환경 복지: 지속 가능성 및 환경 친화성, 사회적 영향, 사회 및 민주주의.
- 책임성: 감사 가능성, 부정적인 영향의 최소화 및 보고, 절충 및 시정.

(미국)NIST(국립 표준기술연구소) AI 위험관리 프레임워크(RMF)

- 2020년 국가 인공지능 이니셔티브법(PL 116-283)은 AI RMF를 개발하도록 함.
- AI RMF는 AI 시스템을 설계, 개발, 배포 또는 사용하는 조직에 *자발적인* 리소스를 제공하여 AI 시스템의 신뢰할 수 있고 책임감 있는 개발 및 사용을 장려하고 다양한 위험을 관리하는 데 도움을 주는 것을 목표로 함.
- 프레임워크는 자발적이고 권리를 보호하며 비부문 특정적이고 사용 사례에 구애받지 않고 모든 규모, 모든 부문, 사회 전체의 조직에 프레임워크의 접근 방식을 구현할 수 있는 유연성을 제공하기 위한 것
- AI RMF는 조직과 개인(AI 행위자)에게 AI 시스템의 신뢰성을 높이는 접근 방식을 제공하도록 설계됨.
- NIST의 이전 정보 위험 관리 및 거버넌스 프레임워크인 2014년에 개발한 사이버보안 프레임워크와 2020년에 개발한 개인정보 보호 프레임워크의 템플릿을 따름.



NIST AI RMF의 구조

- 플레이북
  - AI RMF 출시와 함께 NIST는 "거버넌스, 매핑, 측정 및 관리" 기능과 하위 카테고리에 대한 작업, 참조 및 문서에 대한 추가 제안을 제공하는 GitHub 호스팅 도구인 " 플레이북 "도 출시함.
- 국제 기준과 매핑
  - 이전의 위험관리 프레임워크와 같이 국제 표준과 매핑함. 단 초기 단계이어서 ISO /IEC, OECD, EU AI법 제안, 신뢰할 수 있는 AI에 대한 미국 행정 명령 및 AI 권리 장전의 항목과의 매핑만 있음.
- '사회 기술적 차원'의 위험 관리 - 사회 기술적 차원, 사람과 지구
  - 범용 기술인 AI는 광범위한 기술, 데이터 소스 및 애플리케이션을 포괄하므로, AI의 폭은 정보 기술 위험 관리에 "독특한 어려움"을 야기함. 그래서 AI RMF는 위험 관리 접근 방식에 "사회 기술적" 차원을 도입
  - "사람과 지구"를 고려할 광범위한 결과, 행위자, 이해관계자 및 행위자 전반에 걸쳐 "사회 역학 및 인간 행동"을 포괄하는 넓은 시각을 제공함.
  - AI의 사회 기술적 차원을 AI 시스템 수명주기의 단계 및 관련 행위자와 연결하는 위험을 평가하고 관리하기 위한 일련의 조직 프로세스 및 활동을 제공함.

위험의 프레이밍

위험 측정 위험 --> 허용 범위 설정 --> 위험 우선순위 지정 --> 조직적 통합 및 위험관리

- 위험
  - *위험은* 사건의 발생 가능성과 해당 사건의 결과 규모 또는 정도에 대한 복합적인 척도를 의미, 1) 상황이나 사건이 발생할 경우 발생할 수 있는 부정적인 영향 또는 피해 규모와 2) 발생 가능성의 함수(OMB Circular A- 130:2016).
  - 위험에는 집단의 피해, 민주주의에 미치는 피해도 있음.

본 사안의 경우, 발생 가능한 사건의 부정적인 영향과 발생 가능성 모두 높다고 볼 수 있음,

위험의 측정시 고려할 문제

- 위험이 적절히 평가될 수 있는가?
  - CCTV와 결합된 실시간 원격 생체인식 인공지능 시스템을 수사를 위한 활용에 오용하는 경우 야기되는 문제를 어떻게 평가할 것인지?
- 제3자 데이터와 관련한 위험
  - 외부 CCTV 등 영상자료에 활용할 위험

| 위험 측정 시 겪을 수 있는 문제 |

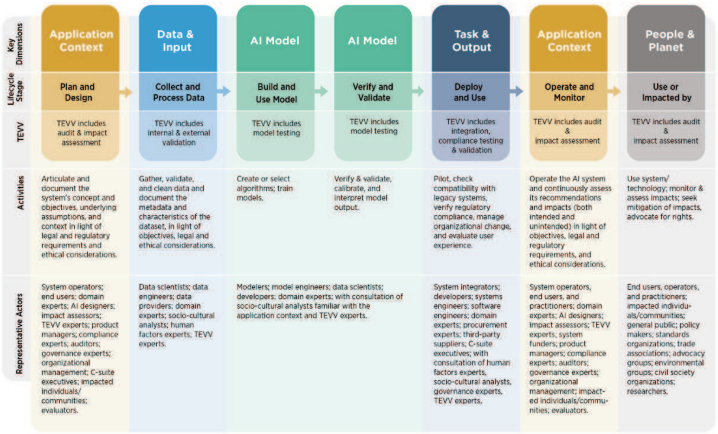
위험 요소	고려 사항
제3자의 데이터, 소프트웨어, 하드웨어	• 내부 거버넌스를 고려해 독립형 또는 통합형으로 AI 시스템을 관리 필요
우발적 위험	• 우발적 위험을 측정하는 기술 개발 필요
신뢰할 수 있는 지표의 가용성	• 모집단에 대한 영향을 측정하는 접근 방식은 피해 요소가 여러 그룹 또는 하위 그룹에 각각 다르게 영향을 미칠 수 있으며 피해를 입은 커뮤니티 또는 하위 그룹이 항상 직접적인 시스템 사용자가 아닐 수 있다는 점을 인식할 때 제대로 작동 가능
AI 단계별 위험	• AI 초기 단계 측정 위험과 각 단계별 위험 상이할 수 있음을 인지
실제 운영 시 위험	• 배포 전 위험과 실제 운영 시 위험이 일치하지 않을 수 있음
불가해성	• AI 시스템의 불투명한 속성(제한된 설명 · 해석), AI 시스템의 개발·배포 시 투명성문서화 부족, AI 시스템에 내재된 불확실성으로 인해 발생할 수 있음
인간 기준선	• 의사결정 등 사람의 활동을 보강하거나 대체하기 위해 기준 설정

| AI 시스템과 관련된 잠재적 피해의 예시 |

사람에 미치는 피해	조직에 미치는 피해	시스템/생태계에 미치는 피해
· 개인: 개인의 자유, 권리, 신체적/심리적 안전 또는 경제적 기회에 미치는 피해 · 집단/커뮤니티: 소수 인종, 민족 집단 차별 등 집단에 미치는 피해 · 사회: 민주적 참여 또는 교육적 접근성에 미치는 피해	· 조직의 비즈니스 운영에 미치는 피해 · 보안 침입 또는 금전적 손실을 통해 조직에 미치는 피해 · 조직 명성에 미치는 피해	· 상호 연결/의존적인 요소 및 리소스에 미치는 피해 · 글로벌 금융 시스템, 공급망 또는 상호 연결 시스템에 미치는 피해 · 천연 자원, 환경 및 지구에 미치는 피해

기존 SW와 비교하여 새롭거나 증가된 AI 관련 위험

- 데이터
  - AI 시스템을 구축하는 데 사용되는 데이터는 AI 시스템의 맥락이나 의도된 사용을 사실이거나 적절하게 표현하지 못할 수 있으며, 실제 정보가 존재하지 않거나 이용 가능하지 않을 수 있음. 유해한 편견 및 기타 데이터 품질 문제는 AI 시스템 신뢰성에 영향을 미쳐 부정적인 영향을 미칠 수 있음.
  - 훈련 작업을 위한 AI 시스템 의존성과 데이터에 대한 의존도가 일반적으로 해당 데이터와 관련된 증가된 양과 복잡성과 결합.
  - AI 시스템을 교육하는 데 사용되는 데이터 세트는 원래 의도한 컨텍스트에서 분리되거나 배포 컨텍스트에 비해 오래되거나 구식이 될 수 있음.
- 훈련 중 의도적이거나 의도하지 않은 변경으로 인해 AI 시스템 성능이 근본적으로 변경될 수 있음.
- 전통적인 소프트웨어 애플리케이션에 수용된 AI 시스템 규모 및 복잡성(시스템에 수십억 또는 심지어 수조 개의 결정 지점이 포함됨).
- 연구를 발전시키고 성능을 향상시킬 수 있는 사전 훈련된 모델을 사용하면 통계적 불확실성 수준이 높아지고 편향 관리, 과학적 타당성 및 재현성에 문제가 발생할 수도 있음.
- 대규모 사전 훈련된 모델의 창발 속성에 대한 실패 모드를 예측하는 데 더 높은 수준의 어려움.
- AI 시스템의 향상된 데이터 집계 기능으로 인한 개인 정보 보호 위험.
- AI 시스템은 데이터, 모델 또는 개념 드리프트로 인해 더 자주 유지 관리하고 수정 유지 관리를 수행하기 위한 트리거가 필요할 수 있음.
- 불투명도가 증가하고 재현성에 대한 우려가 높아짐.
- 소프트웨어 테스트 표준이 덜 개발되어 있고 가장 단순한 경우를 제외하고는 전통적으로 엔지니어링된 소프트웨어에 기대되는 표준에 따라 AI 기반 사례를 문서화할 수 없음.
- AI 시스템은 기존 코드 개발과 동일한 제어 대상이 아니기 때문에 정기적인 AI 기반 소프트웨어 테스트를 수행하거나 무엇을 테스트할지 결정하는 데 어려움이 있음.
- AI 시스템 개발을 위한 계산 비용과 환경 및 지구에 미치는 영향.
- 통계적 측정 이상으로 AI 기반 시스템의 부작용을 예측하거나 감지할 수 없음



AI 수명주기 단계 전반의 AI 행위자.(NIST RMF)

중심에 있는 사람과 지구 차원은 인권과 사회 및 지구의 광범위한 복지를 나타냄(NIST RMF)

위험 관리 프레임워크의 행위자(관계자)

AI 위험과 신뢰성

- 신뢰할 수 있는 AI 시스템의 특징은 **유효하고 신뢰할 수 있으며, 안전하고, 보안성이 있고, 탄력적이며, 책임감 있고 투명하고, 설명 가능하고 해석 가능하며, 개인 정보 보호가 강화되고, 유해한 편견이 관리되어 공정함.**
- 상충관계
  - 신뢰성 특성은 **서로 영향**. 매우 안전하지만 불공평한 시스템, 정확하지만 불투명하고 해석하기 어려운 시스템, 부정확하지만 안전하고 개인정보 보호가 보호되고 투명한 시스템.
  - **궁극적으로 신뢰는 다양한 스펙트럼에 걸쳐 있으며 가장 약한 특성만큼만 강한 사회적 개념임.**
  - 상충시 판단
    - 판단은 신뢰성 특성과 상대적인 위험, 영향, 비용 및 이점에 대한 상황별 평가를 기반으로 하고 광범위한 이해 당사자의 정보를 바탕으로 이루어져야 함.

신뢰성 특성은 서로 영향

매우 안전하지만 불공평한 시스템, 정확하지만 불투명하고 해석하기 어려운 시스템, 부정확하지만 안전하고 개인정보 보호가 강화되고 투명한 시스템 등. 이들은 모두 바람직하지 않음.

위험 관리에 대한 포괄적인 접근 방식에서는 신뢰성 특성 간의 균형을 맞추는 것이 필요

AI 기술이 주어진 상황이나 목적에 적합하거나 필요한 도구인지, 그리고 이를 책임감 있게 사용하는 방법을 결정하는 것은 모든 AI 행위자의 공동 책임

AI 시스템을 위탁하거나 배포하기로 한 결정은 신뢰성 특성과 상대적인 위험, 영향, 비용 및 이점에 대한 상황별 평가를 기반으로 하고 광범위한 이해 당사자의 정보를 바탕으로 이루어져야 함."(NIST AI RMF)

정확성, 견고성, 유효성, 타당성

- **Validation 유효성**은 "객관적인 증거 제공을 통해 특정 의도된 용도 또는 적용에 대한 요구 사항이 충족되었음을 확인하는 것"(ISO 9000:2015).
- **신뢰성(Reliability)**은 동일한 표준에서 "주어진 조건 하에서 **주어진 시간 간격 동안 고장 없이 요구되는 대로 수행하는 품목의 능력**"(ISO/IEC TS 5723:2022).
- **정확성(Accuracy)**은 "**관찰, 계산 또는 추정 결과가 참값 또는 참으로 인정된 값에 근접한 정도**"(ISO/IEC TS 5723:2022)
- **견고성 또는 일반화 가능성(Robustness or generalizability)**은 "**다양한 상황에서 성능 수준을 유지하는 시스템의 능력**"(ISO/IEC TS 5723:2022). **견고성은 시스템이 예상된 용도에서와 똑같이 작동할 뿐만 아니라 예상치 못한 환경에서 작동할 경우 사람에게 미칠 수 있는 피해를 최소화하는 방식으로 작동해야.**
- 배포된 AI 시스템의 **유효성과 신뢰성(Validity and reliability)**은 **시스템이 의도한 대로 작동하는지 확인하는 지속적인 테스트 또는 모니터링을 통해 평가.**
- **타당성, 정확성, 견고성 및 신뢰성 측정은 신뢰성(trustworthiness)에 기여하며 특정 유형의 실패가 더 큰 해를 끼칠 수 있다는 점을 고려해야.**
- **AI 위험 관리 노력은 잠재적인 부정적인 영향을 최소화하는 것을 우선시해야 하며 AI 시스템이 오류를 감지하거나 수정할 수 없는 경우 인간의 개입을 포함해야** 할 수도.

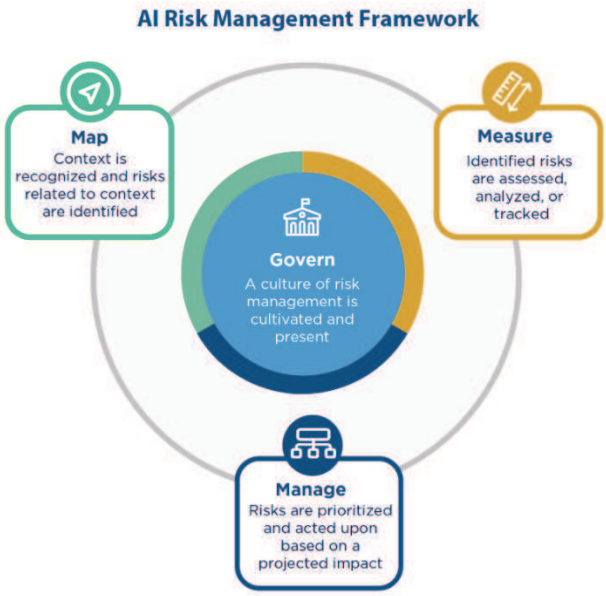
안전성

- AI 시스템은 "정의된 조건에서 인간의 생명, 건강, 재산 또는 환경이 위험에 빠지는 상태로 이어져서는 안 됨"(ISO/IEC TS 5723:2022). AI 시스템의 안전한 작동은 다음을 통해 개선됨.
  - 책임 있는 설계, 개발 및 배포 방식;
  - 시스템의 책임 있는 사용에 관해 배포자에게 명확한 정보 제공.
  - 배포자와 최종 사용자의 책임 있는 의사 결정
  - 사건의 경험적 증거를 바탕으로 위험에 대한 설명 및 문서화.
- 다양한 유형의 안전 위험에는 제시된 잠재적 위험의 심각도와 상황에 따라 맞춤형 AI 위험 관리 접근 방식이 필요.
- 심각한 부상이나 사망의 잠재적 위험을 초래하는 안전 위험은 가장 긴급한 우선순위 지정과 가장 철저한 위험 관리 프로세스를 요구.
- 수명주기 동안 안전을 고려하고 계획 및 설계를 가능한 한 빨리 시작하면 시스템을 위험하게 만들 수 있는 오류나 조건을 방지할 수 있음

책임성, 투명함

- 책임과 투명성
  - 신뢰할 수 있는 AI는 책임에 달려 있으며, 책임은 투명성을 전제로 함
- 투명성
  - AI 라이프사이클 단계를 기반으로 하고 AI 시스템과 상호작용하거나 AI 시스템을 사용하는 AI 행위자 또는 개인의 역할이나 지식에 맞춰진 적절한 수준의 정보에 대한 액세스를 제공.
  - 투명성은 더 높은 수준의 이해를 촉진함으로써 AI 시스템에 대한 신뢰도를 높임.
- 책임성과 투명성
  - 설계 결정 및 교육 데이터부터 모델 교육, 모델의 구조, 의도된 사용 사례, 배포, 배포 후 또는 최종 사용자 결정이 내려진 방법과 시기 및 주체에 이르기까지 다양
  - 부정확하거나 불이익 또는 부정적인 영향을 초래하는 AI 시스템 출력에는 투명성이 필요한 경우가 많음
- 고려사항
  - 심각한 위험과 결과에는 투명성과 책임을 적극적으로 조정. 필요한 자원 수준과 독점 정보 보호 필요성을 포함하여 그 영향 고려.
  - AI 시스템 및 관련 문서를 위한 투명성 도구가 계속 발전함에 따라 AI 시스템 개발자는 AI 시스템이 의도한 대로 사용되는지 확인하기 위해 AI 배포자와 협력하여 다양한 유형의 투명성 도구를 테스트하도록 권장

미국 NIST AI RMF(Risk Management Framework)의 위험 관리



카테고리	하위 카테고리
1 AI 위험 매핑, 측정 및 관리와 관련된 조직 전반의 정책, 프로세스, 절차 및 수행 기준이 투명하고 효과적으로 구현된다.	**1.1** : AI와 관련된 **법률 및 규제 요구 사항**을 이해하고 관리하며 문서화합니다.
	1.2 : 신뢰할 수 있는 AI의 특성은 **조직의 정책, 프로세스, 절차 및 관행에 통합**됩니다.
	1.3 : 조직의 **위험 허용 범위를 기반으로 필요한 위험 관리 활동 수준을 결정하기 위한 프로세스, 절차 및 관행**이 마련되어 있습니다.
	1.4 : 위험 관리 프로세스와 그 결과는 **조직의 위험 우선순위**에 따라 **투명한 정책, 절차 및 기타 통제**를 통해 확립됩니다.
	1.5 : 위험 관리 프로세스와 그 결과에 대한 **지속적인 모니터링**과 **정기 검토**가 계획되고 정기 검토 **빈도** 결정을 포함하여 조직의 역할과 책임이 명확하게 정의됩니다.
	1.6 : AI 시스템의 목록을 작성하는 메커니즘이 마련되어 있으며 조직의 위험 우선순위에 따라 자원이 배정됩니다.
	1.7 : 위험을 증가시키거나 조직의 신뢰성을 저하시키지 않는 방식으로 AI 시스템을 **안전하게 폐기하고 단계적으로 폐지**하기 위한 프로세스와 절차가 마련되어 있습니다.
2 : 적절한 팀과 개인이 AI 위험을 매핑, 측정 및 관리하기 위한 권한을 부여받고 책임을 지며 교육을 받을 수 있도록 책임 구조가 마련되어 있습니다.	**2.1** : AI 위험 매핑, 측정, 관리와 관련된 **역할과 책임, 커뮤니케이션 라인**이 **문서화되어 조직 전체의 개인과 팀에 명확**합니다.
	2.2 : **조직의 직원과 파트너**는 AI 위험 관리 교육을 받아 관련 정책, 절차 및 계약에 따라 직무와 책임을 수행할 수 있습니다.
	2.3 : 조직의 **경영진**은 AI 시스템 개발 및 배포와 관련된 위험에 대한 결정을 책임집니다.

거버넌스

카테고리	하위 카테고리
3 : 수명주기 전반에 걸쳐 AI 위험을 매핑, 측정 및 관리할 때 **인력 다양성, 형평성, 포용성 및 접근성** 프로세스가 우선시됩니다.	3.1 : 수명주기 전반에 걸쳐 AI 위험을 매핑, 측정 및 관리하는 것과 관련된 의사결정은 **다양한 팀(예: 다양한 인구통계, 분야, 경험, 전문 지식 및 배경)을 통해** 이루어집니다.
	3.2 : **인간-AI 구성 및 AI 시스템 감독**에 대한 역할과 책임을 정의하고 **차별화하기 위한** 정책과 절차가 마련되어 있습니다.
4: 조직 팀은 AI 위험을 고려하고 **전달하는 문화에 전념**합니다.	4.1 : **잠재적인 부정적 영향을 최소화**하기 위해 AI 시스템의 설계, 개발, 배포 및 사용에 있어 **비판적 사고와 안전 우선 사고 방식을 육성하기 위한 조직 정책 및 관행**이 마련되어 있습니다.
	4.2 : 조직 팀은 자신이 설계, 개발, 배포, 평가, 사용하는 AI 기술의 **위험과 잠재적 영향을 문서화**하고 그 영향에 대해 보다 **광범위하게 소통**합니다.
	4.3 : **AI 테스트, 사고 식별, 정보 공유를 가능하게 하는 조직적 관행**이 마련되어 있습니다.
5 : 관련 AI 행위자와의 **강력한 참여**를 위한 프로세스가 마련되어 있습니다.	5.1 : AI 위험과 관련된 잠재적인 개인 및 사회적 영향에 관해 AI 시스템을 개발하거나 배포한 **팀 외부의 피드백을 수집, 고려, 우선순위 지정 및 통합하기 위한 조직 정책 및 관행**이 마련되어 있습니다.
	5.2 : AI 시스템을 개발하거나 배포한 팀이 관련 AI 행위자로부터 **판정된 피드백을 시스템 설계 및 구현에 정기적으로 통합할 수 있도록 하는 메커니즘**이 확립되었습니다.
6 : **제3자 소프트웨어와 데이터, 기타 공급망 문제로 인해 발생하는 AI 위험**과 이점을 해결하기 위한 정책과 절차가 마련되어 있습니다.	6.1 : 제3자의 지적 재산권이나 기타 권리 침해 위험을 포함하여 **제3자 단체와 관련된 AI 위험**을 해결하는 정책 및 절차가 마련되어 있습니다.
	6.2 : 위험도가 높은 것으로 간주되는 **제3자 데이터 또는 AI 시스템의 오류나 사고를 처리**하기 위한 비상 프로세스가 마련되어 있습니다.

거버넌스

실시간 CCTV 생체인식 AI 시스템과 거버넌스 관련 검토

항목	검토사항
1.1 : AI와 관련된 **법률 및 규제 요구 사항**을 이해하고 관리하며 문서화.	법적 근거가 미비한 상태에서 해당 AI 시스템의 신뢰성 인정은 난점이 있음.
1.2 : 신뢰할 수 있는 AI의 특성은 **조직의 정책, 프로세스, 절차 및 관행에 통합**.	신뢰성이 정책, 프로세스, 절차, 관행에 모두 관철되어 있는지를 검토해야 함.
3.1 : 수명주기 전반에 걸쳐 AI 위험을 매핑, 측정 및 관리하는 것과 관련된 의사결정은 다양한 팀(예: 다양한 인구통계, 분야, 경험, 전문 지식 및 배경)을 통해 이루어집니다.	위험 의사결정 다양성
3.2 : 인간-AI 구성 및 AI 시스템 감독에 대한 역할과 책임을 정의하고 차별화하기 위한 정책과 절차가 마련되어 있습니다.	인간의 결정
4.1 : **잠재적인 부정적 영향을 최소화**하기 위해 AI 시스템의 설계, 개발, 배포 및 사용에 있어 **비판적 사고와 안전 우선 사고방식을 육성하기 위한 조직 정책 및 관행**이 마련되어 있습니다.	비판적 사고와 안전 우선 사고방식
4.2 : 조직 팀은 자신이 설계, 개발, 배포, 평가, 사용하는 AI 기술의 **위험과 잠재적 영향을 문서화**하고 그 영향에 대해 보다 **광범위하게 소통**합니다.	위험의 잠재적 영향, 소통
4.3 : **AI 테스트, 사고 식별, 정보 공유를 가능하게 하는 조직적 관행**이 마련되어 있습니다.	사고 정보 공유 조직적 관행
5.1 : AI 위험과 관련된 잠재적인 개인 및 사회적 영향에 관해 AI 시스템을 개발하거나 배포한 **팀 외부의 피드백을 수집, 고려, 우선순위 지정 및 통합하기 위한 조직 정책 및 관행**이 마련되어 있습니다.	외부 피드백
5.2 : AI 시스템을 개발하거나 배포한 팀이 관련 AI 행위자로부터 **판정된 피드백을 시스템 설계 및 구현에 정기적으로 통합할 수 있도록 하는 메커니즘**이 확립되었습니다.	외부 피드백 설계에 구현
6.1 : 제3자의 지적 재산권이나 기타 권리 침해 위험을 포함하여 **제3자 단체와 관련된 AI 위험**을 해결하는 정책 및 절차가 마련되어 있습니다.	제3자 관련 위험
6.2 : 위험도가 높은 것으로 간주되는 **제3자 데이터 또는 AI 시스템의 오류나 사고를 처리**하기 위한 비상 프로세스가 마련되어 있습니다.	제3자 데이터 등

매핑

카테고리	하위 카테고리
매핑 1 : 맥락이 확립되고 이해됩니다.	**매핑 1.1** : AI 시스템이 배포될 **의도된 목적, 잠재적으로 유익한 용도, 상황별 법률, 규범 및 기대, 예상 설정을 이해하고 문서화**합니다. 고려 사항에는 다음이 포함됩니다: 특정 사용자 집합 또는 유형과 기대치; **개인, 지역 사회, 조직, 사회 및 지구에 대한 시스템 사용의 잠재적인 긍정적 및 부정적 영향**; 개발 또는 제품 AI 수명주기 전반에 걸쳐 AI 시스템 목적, 사용 및 위험에 대한 가정 및 관련 제한사항 관련 TEVV 및 시스템 측정항목.
	매핑 1.2 : 맥락을 구축하기 위한 **학제간 AI 행위자, 역량, 기술 및 역량은 인구통계학적 다양성과 광범위한 영역 및 사용자 경험 전문 지식을 반영하며 이들의 참여가 문서화**됩니다. **학제 간 협력 기회**가 우선시됩니다.
	매핑 1.3 : AI 기술에 대한 조직의 사명과 관련 목표를 이해하고 문서화합니다.
	매핑 1.4 : 비즈니스 가치 또는 비즈니스 사용 맥락이 명확하게 정의되었거나 기존 AI 시스템을 평가하는 경우 재평가되었습니다.
	매핑 1.5 : **조직의 위험 허용 범위가 결정되고 문서화**됩니다.
	매핑 1.6 : 시스템 요구 사항(예: "시스템은 사용자의 개인 정보를 존중해야 합니다")은 **관련 AI 행위자로부터 도출되고 이해됩니다. 설계 결정은 AI 위험을 해결하기 위해 사회 기술적 영향을 고려**합니다.
MAP 2 : AI 시스템의 분류가 수행됩니다.	**매핑 2.1** : AI 시스템이 지원할 작업을 구현하는 데 사용되는 특정 작업과 방법이 정의됩니다(예: 분류자, 생성 모델, 추천자).
	매핑 2.2 : **AI 시스템의 지식 한계와 시스템 출력을 인간이 활용하고 감독하는 방법에 대한 정보가 문서화**되어 있습니다. **문서는 관련 AI 행위자가 결정을 내리고 후속 조치를 취할 때 도움**이 되는 **충분한 정보를 제공**합니다.
	매핑 2.3 : **실험 설계, 데이터 수집 및 선택(예: 가용성, 대표성, 적합성), 시스템 신뢰성 및 구성 검증과 관련된 사항을 포함하여 과학적 무결성 및 TEVV 고려 사항이 식별되고 문서화**됩니다.

매핑

카테고리	하위 카테고리
매핑 3 : 적절한 벤치마크와 비교하여 AI 기능, 목표 사용법, 목표, 예상 이점 및 비용을 이해합니다.	**매핑 3.1** : 의도된 AI 시스템 기능 및 성능의 잠재적 이점을 조사하고 문서화합니다.
	매핑 3.2 : 예상되거나 실현된 **AI 오류나 시스템 기능 및 신뢰성(조직의 위험 허용 범위와 관련됨)으로 인해 발생하는 비금전적 비용을 포함한 잠재적 비용**을 조사하고 **문서화**합니다.
	매핑 3.3 : 시스템 성능, 확립된 컨텍스트, AI 시스템 분류를 기반으로 **대상 애플리케이션 범위를 지정하고 문서화**합니다.
	매핑 3.4 : AI 시스템 성능 및 신뢰성, 관련 기술 표준 및 인증에 대한 **운영자 및 실무자의 숙련도를 위한 프로세스**가 정의, 평가 및 문서화됩니다.
	매핑 3.5 : 사람의 감독을 위한 프로세스는 관리 기능 의 조직 정책에 따라 정의, 평가 및 문서화됩니다 .
매핑 4 : **타사 소프트웨어 및 데이터를 포함한 AI 시스템의 모든 구성 요소에 대한 위험과 이점**이 매핑되어 있습니다.	**매핑 4.1** : 제3자의 지적 재산권 또는 기타 권리 침해 위험과 마찬가지로 **제3자 데이터 또는 소프트웨어 사용**을 포함하여 AI 기술 및 해당 구성 요소의 법적 위험을 매핑하기 위한 접근 방식이 확립되어 있고, 준수되고, 문서화되어 있습니다.
	매핑 4.2 : 타사 AI 기술을 포함한 AI 시스템 구성 요소에 대한 내부 위험 제어가 식별되고 문서화됩니다.
매핑 5 : **개인, 집단, 지역 사회, 조직, 사회에 미치는 영향이 특성화**되어 있습니다.	**매핑 5.1** : **예상되는 사용, 유사한 맥락에서 AI 시스템의 과거 사용, 공공 사건 보고, AI 시스템을 개발하거나 배포한 팀 외부 사람들의 피드백을 기반으로 확인된 각 영향(잠재적으로 유익하거나 유해함)의 가능성과 규모, 또는 기타 데이터가 식별되고 문서화**됩니다.
	매핑 5.2 : **관련 AI 행위자와의 정기적인 참여를 지원하고 긍정적, 부정적, 예상치 못한 영향에 대한 피드백을 통합하기 위한 관행과 인력이 마련되어 문서화**되어 있습니다.

실시간 CCTV 생체인식 AI 시스템과 매핑 관련 검토

항목	검토사항
1.1 : AI 시스템이 배포될 **의도된 목적, 잠재적으로 유익한 용도, 상황별 법률, 규범 및 기대, 예상 설정을 이해하고 문서화**합니다. 고려 사항에는 다음이 포함됩니다: 특정 사용자 집합 또는 유형과 기대치; **개인, 지역 사회, 조직, 사회 및 지구에 대한 시스템 사용의 잠재적인 긍정적 및 부정적 영향**; 개발 또는 제품 AI 수명주기 전반에 걸쳐 AI 시스템 목적, 사용 및 위험에 대한 가정 및 관련 제한사항 관련 TEVV 및 시스템 측정항목.	지역사회 등에 대한 잠재적 영향 문서화
1.2 : 맥락을 구축하기 위한 **학제간 AI 행위자, 역량, 기술 및 역량은 인구통계학적 다양성과 광범위한 영역 및 사용자 경험 전문 지식을 반영하며 이들의 참여가 문서화**됩니다. **학제 간 협력 기회**가 우선시됩니다.	다양한 행위자의 참여 문서화
1.5 : **조직의 위험 허용 범위가 결정되고 문서화**됩니다.	위험허용범위 문서화
1.6 : 시스템 요구 사항(예: "시스템은 사용자의 개인 정보를 존중해야 합니다")은 **관련 AI 행위자로부터 도출되고 이해됩니다. 설계 결정은 AI 위험을 해결하기 위해 사회 기술적 영향을 고려**합니다.	관련 행위자로부터 도출
2.2 : **AI 시스템의 지식 한계와 시스템 출력을 인간이 활용하고 감독하는 방법에 대한 정보가 문서화**되어 있습니다. **문서는 관련 AI 행위자가 결정을 내리고 후속 조치를 취할 때 도움**이 되는 **충분한 정보를 제공**합니다.	시스템 한계 정보 문서화, 충분한 정보
2.3 : **실험 설계, 데이터 수집 및 선택(예: 가용성, 대표성, 적합성), 시스템 신뢰성 및 구성 검증과 관련된 사항을 포함하여 과학적 무결성 및 TEVV 고려 사항이 식별되고 문서화**됩니다.	과학적 무결성 등 문서화
3.2 : 예상되거나 실현된 **AI 오류나 시스템 기능 및 신뢰성(조직의 위험 허용 범위와 관련됨)으로 인해 발생하는 비금전적 비용을 포함한 잠재적 비용**을 조사하고 **문서화**합니다.	오류로 인한 비용 문서화
3.4 : AI 시스템 성능 및 신뢰성, 관련 기술 표준 및 인증에 대한 **운영자 및 실무자의 숙련도를 위한 프로세스**가 정의, 평가 및 문서화.	실무자 숙련도를 위한 프로세스 문서화
3.5 : 사람의 감독을 위한 프로세스는 관리 기능 의 조직 정책에 따라 정의, 평가 및 문서화됩니다 .	사람의 감도 프로세스 문서화
4.1 : 제3자의 지적 재산권 또는 기타 권리 침해 위험과 마찬가지로 **제3자 데이터 또는 소프트웨어 사용**을 포함하여 AI 기술 및 해당 구성 요소의 법적 위험을 매핑하기 위한 접근 방식이 확립되어 있고, 준수되고, 문서화되어 있습니다.	제3자 데이터 위험 매핑 문서화
5.1 : **예상되는 사용, 유사한 맥락에서 AI 시스템의 과거 사용, 공공 사건 보고, AI 시스템을 개발하거나 배포한 팀 외부 사람들의 피드백을 기반으로 확인된 각 영향(잠재적으로 유익하거나 유해함)의 가능성과 규모, 또는 기타 데이터가 식별되고 문서화**됩니다.	과거, 공공 정보 문서화
5.2 : **관련 AI 행위자와의 정기적인 참여를 지원하고 긍정적, 부정적, 예상치 못한 영향에 대한 피드백을 통합하기 위한 관행과 인력이 마련되어 문서화**되어 있습니다.	피드백, 참여 통합

측정

카테고리	하위 카테고리
측정 1 적절한 방법 및 지표를 식별하고 적용한다.	1.1 가장 중요한 AI 위험을 우선적으로 구현하기 위해 매핑 기능을 통해 열거된 **AI 위험 측정 방법 및 지표**를 선택한다. **측정하지 않거나 측정할 수 없는 위험 또는 신뢰도 특성을 적절히 문서화**한다.
	1.2 **오류 보고서 및 커뮤니티에 대한 잠재적 영향을 포함하여 AI 지표의 적절성 및 기존 제어의 효율성을 정기적으로 평가 및 업데이트**한다.
	1.3 시스템의 일선 개발자 또는 독립 평가자의 역할을 하지 않은 **내부 전문가를 정기적 평가 및 업데이트에 참여**시킨다. **도메인 전문가, 사용자, AI 시스템을 개발 또는 배포한 팀의 외부 AI 행위자 및 영향을 받는 커뮤니티**는 조직의 위험 허용 범위에 따라 필요한 평가를 지원한다.
측정 2 신뢰할 수 있는 특성에 대해 AI 시스템을 평가한다.	2.1 TEVV 중에 사용된 도구의 테스트 세트, 지표 및 세부 정보를 **문서화**한다.
	2.2 **인간 피실험자와 관련된 평가는 관련 요구 사항(인간 피실험자 보호 포함)을 충족하고 모집단을 대표**한다.
	2.3 AI 시스템의 **성능 또는 보증 기준을 정성적 또는 정량적으로 측정하고 배포 조건과 유사한 조건에서 입증**한다. **조치를 문서화**한다.
	2.4 매핑 기능에서 식별된 AI 시스템 및 구성 요소의 기능과 동작은 **제조시 모니터**링된다.
	2.5 배포할 AI 시스템이 **타당하고 신뢰할 수 있는지를 입증**한다. **기술 개발 조건 이외의 일반화 한계를 문서화**한다.
	2.6 **매핑 기능에서 식별되는 안전 위험에 대해 AI 시스템을 정기적으로 평가**한다. 배포할 AI 시스템이 안전하다는 것을 입증하고 **남은 부정적 위험은 위험 허용 범위를 초과하지 않아야** 한다. AI 시스템이 정보 한계를 넘어 작동하도록 구성된 경우 안전에 실패할 수 있다. **안전 지표는 시스템의 신뢰성, 견고성, 실시간 모니터링 및 AI 시스템 오류에 대한 응답 시간을 반영**한다.
	2.7 매핑 기능에서 식별된 AI 시스템의 **보안 및 탄력성**을 평가 및 문서화한다.
	2.8 매핑 기능에서 식별된 **투명성 및 책임과 관련된 위험을 조사하고 문서화**한다.

측정

카테고리	하위 카테고리
	2.9 AI 모델을 **설명, 검증 및 문서화해야 하며 책임 있는 사용과 기능에 대해 알리기 위해 AI 시스템 결과를 매핑 기능을 통해 식별한 상황 내에서 해석**해야 한다.
	2.10 매핑 기능에서 식별된 AI 시스템의 **개인정보보호 위험을 조사하고 문서화**한다.
	2.11 매핑 기능에서 식별된 **공정성 및 편향을 평가하고 그 결과를 문서화**한다.
	2.12 매핑 기능에서 식별된 AI 모델 훈련 및 관리 활동에 대한 환경적 영향 및 지속 가능성을 평가하고 문서화한다.
	2.13 측정 기능에서 사용된 **TEVV 지표 및 프로세스의 효율성을 평가하고 문서화**한다.
측정 3 AI 위험을 시간 경과에 따라 추적하는 메커니즘을 구축한다.	3.1 배포 상황 내에서 잠재적/실제적 성능 등의 요소를 기반으로 기존의, **예상치 못한, 새로운 AI 위험을 정기적으로 식별하고 추적하기 위한 접근 방법, 인력 및 문서를 구축**한다.
	3.2 현재 가용 측정 기술을 사용하여 **AI 위험을 평가하기 어렵거나 관련 지표를 아직 사용할 수 없는 경우 위험 추적 접근 방법**이 고려된다.
	3.3 문제를 보고하고 시스템 결과에 **이의를 제기하기 위한 최종 사용자 및 영향을 받는 커뮤니티의 피드백 프로세스**를 구축하여 AI 시스템 평가 지표에 통합한다.
측정 4 측정 효율성에 대한 피드백을 수집하고 평가한다.	4.1 AI 위험을 식별하기 위한 측정 방법을 배포 상황과 연관시켜 **도메인 전문가 및 기타 최종 사용자와의 협의를 통해 정보**를 얻는다. 접근 방법을 문서화한다.
	4.2 시스템이 의도한 바에 따라 일관되게 수행되는지를 검증하기 위해 **도메인 전문가 및 관련 AI 행위자를 통해 배포 상황 및 AI 주기 전반에 걸친 AI 시스템 신뢰도에 대한 측정 결과**를 얻는다. 결과를 문서화한다.
	4.3 커뮤니티 및 관련 AI 행위자와의 협의를 기반으로 측정한 성능의 개선 또는 감소, 상황과 관련된 위험 및 신뢰도 특성에 관한 **현장 데이터를 식별하고 문서화**한다.

실시간 CCTV 생체인식 AI 시스템과 측정 관련검토

항목	검토할 사항
1.1 가장 중요한 AI 위험을 우선적으로 구현하기 위해 매핑 기능을 통해 열거된 **AI 위험 측정 방법 및 지표**를 선택한다. **측정하지 않거나 측정할 수 없는 위험 또는 신뢰도 특성을 적절히 문서화**한다.	측정할 수 없는 위험 문서화
1.2 **오류 보고서 및 커뮤니티에 대한 잠재적 영향을 포함하여 AI 지표의 적절성 및 기존 제어의 효율성을 정기적으로 평가 및 업데이트**한다.	커뮤니티에 대한 영향 포함 정기 평가
1.3 시스템의 일선 개발자 또는 독립 평가자의 역할을 하지 않은 **내부 전문가를 정기적 평가 및 업데이트에 참여**시킨다. **도메인 전문가, 사용자, AI 시스템을 개발 또는 배포한 팀의 외부 AI 행위자 및 영향을 받는 커뮤니티**는 조직의 위험 허용 범위에 따라 필요한 평가를 지원한다.	비관여자, 외부자 평가 정기적 참여
2.1 TEVV 중에 사용된 도구의 테스트 세트, 지표 및 세부 정보를 **문서화**한다.	TEVV 문서화
2.2 **인간 피실험자와 관련된 평가는 관련 요구 사항(인간 피실험자 보호 포함)을 충족하고 모집단을 대표**한다.	모집단 대표성
2.3 AI 시스템의 **성능 또는 보증 기준을 정성적 또는 정량적으로 측정하고 배포 조건과 유사한 조건에서 입증**한다. **조치를 문서화**한다.	배포조건에서의 성능 문서화
2.5 배포할 AI 시스템이 **타당하고 신뢰할 수 있는지를 입증**한다. **기술 개발 조건 이외의 일반화 한계를 문서화**한다.	배포시의 타당성, 일반화 한계 문서화
2.6 **매핑 기능에서 식별되는 안전 위험에 대해 AI 시스템을 정기적으로 평가**한다. 배포할 AI 시스템이 안전하다는 것을 입증하고 **남은 부정적 위험은 위험 허용 범위를 초과하지 않아야** 한다. AI 시스템이 정보 한계를 넘어 작동하도록 구성된 경우 안전에 실패할 수 있다. **안전 지표는 시스템의 신뢰성, 견고성, 실시간 모니터링 및 AI 시스템 오류에 대한 응답 시간을 반영**한다.	남은 위험의 범위
2.7 매핑 기능에서 식별된 AI 시스템의 **보안 및 탄력성**을 평가 및 문서화한다.	보안과 탄력성 문서화
2.8 매핑 기능에서 식별된 **투명성 및 책임과 관련된 위험을 조사하고 문서화**한다.	투명성, 책임 문서화

실시간 CCTV 생체인식 AI 시스템과 측정 관련 검토

항목	검토할 사항
2.9 AI 모델을 **설명, 검증 및 문서화해야 하며 책임 있는 사용과 기능에 대해 알리기 위해 AI 시스템 결과를 매핑 기능을 통해 식별한 상황 내에서 해석**해야 한다.	상황 내에서 책임있는 사용 문서화
2.10 매핑 기능에서 식별된 AI 시스템의 **개인정보보호 위험을 조사하고 문서화**한다.	개인정보보호 위험 문서화
2.11 매핑 기능에서 식별된 **공정성 및 편향을 평가하고 그 결과를 문서화**한다.	공정성과 편향 평가와 문서화
2.13 측정 기능에서 사용된 **TEVV 지표 및 프로세스의 효율성을 평가하고 문서화**한다.	TEVV 효율성 평가 문서화
3.1 배포 상황 내에서 잠재적/실제적 성능 등의 요소를 기반으로 기존의, **예상치 못한, 새로운 AI 위험을 정기적으로 식별하고 추적하기 위한 접근 방법, 인력 및 문서화**한다.	새로운 AI 위험 식별 접근 방법, 문서, 인력
3.2 현재 가용 측정 기술을 사용하여 **AI 위험을 평가하기 어렵거나 관련 지표를 아직 사용할 수 없는 경우 위험 추적 접근 방법**이 고려된다.	평가하기 어려운 위험 추적
3.3 문제를 보고하고 시스템 결과에 **이의를 제기하기 위한 최종 사용자 및 영향을 받는 커뮤니티의 피드백 프로세스**를 구축하여 AI 시스템 평가 지표에 통합한다.	이의 제기 최종사용자 피드백 프로세스
4.1 AI 위험을 식별하기 위한 측정 방법을 배포 상황과 연관시켜 **도메인 전문가 및 기타 최종 사용자와의 협의를 통해 정보**를 얻는다. 접근 방법을 문서화한다.	배포 상황의 최종사용자 위험 협의
4.2 시스템이 의도한 바에 따라 일관되게 수행되는지를 검증하기 위해 **도메인 전문가 및 관련 AI 행위자를 통해 배포 상황 및 AI 주기 전반에 걸친 AI 시스템 신뢰도에 대한 측정 결과**를 얻는다. 결과를 문서화한다.	행위자를 통한 신뢰도 측정 문서화
4.3 커뮤니티 및 관련 AI 행위자와의 협의를 기반으로 측정한 성능의 개선 또는 감소, 상황과 관련된 위험 및 신뢰도 특성에 관한 **현장 데이터를 식별하고 문서화**한다.	현장 데이터 문서화

관리

카테고리	하위 카테고리
관리 1 매핑 및 측정 기능으로부터 얻은 **평가 및 기타 분석 결과를 기반으로 AI 위험에 대해 우선순위 부여, 대응하며, 관리**한다.	1.1 AI 시스템이 의도한 목적 및 **목표를 달성했는지 여부와 시스템의 개발 또는 배포를 진행해야 하는지 여부에 대한 결정**을 내린다.
	1.2 문서화된 **AI 위험은 영향, 가능성, 가용 리소스 또는 방법에 따라 그 우선 순위가 지정**된다.
	1.3 매핑 기능을 통해 식별된 **우선 순위가 높은 AI 위험에 대응하기 위한 방법을 개발, 계획 및 문서화**한다. 위험 대응 옵션에는 완화, 이전, 회피 또는 수용이 포함된다.
	1.4 AI 시스템의 **후속 취득자 및 최종 사용자 모두에 대한 부정적인 잔류 위험(완화되지 않은 모든 위험의 합계로 정의됨)을 문서화**한다.
관리 2 **관련 AI 행위자의 개입을 통해 AI 이점을 극대화하고 부정적인 영향을 최소화하기 위한 전략을 계획, 준비, 구현, 문서화하고 해당 정보를 제공**한다.	2.1 **잠재적 영향의 규모 또는 가능성을 줄이기 위해 실행 가능한 비-AI 대체 시스템, 접근 방식 또는 방법과 함께 AI 위험을 관리하는 데 필요한 리소스를 고려**한다.
	2.2 배포된 AI 시스템의 가치를 유지하기 위한 메커니즘을 구축하고 적용한다.
	2.3 **이전에 알려지지 않은 위험이 식별될 경우 해당 위험에 대응하고 그로부터 복구하기 위한 절차**를 준수한다.
	2.4 의도한 목적과는 다른 성능 또는 결과를 나타내는 **AI 시스템을 대체, 해제 또는 비활성화하기 위한 메커니즘을 마련하고 관련 책무를 할당하고 파악**한다.
관리 3 제3자 기관의 AI 위험 및 이점을 관리한다.	3.1 **제3자 리소스의 AI 위험 및 이점을 정기적으로 모니터**링하고 위험 제어를 적용하고 문서화한다.
	3.2 **AI 시스템의 정기적 모니터링 및 유지 관리의 일환**으로 개발용으로 사용되는 사전 학습된 모델을 모니터링한다.
관리 4 식별 및 측정된 **AI 위험에 대해 위험 처리(대응 및 복구 포함) 및 커뮤니케이션 계획을 문서화하고 이를 정기적으로 모니터링**한다.	4.1 배포 후 AI 시스템에 대한 **모니터링 계획을 구현**한다. 여기에는 사용자 및 기타 **관련 AI 행위자의 의견을 수집하고 평가하기 위한 메커니즘, 이의 제기, 중단, 해제, 사고 대응, 복구 및 변경 관리가 포함**된다.
	4.2 지속적인 개선 활동이 AI 시스템 업데이트에 통합되며, 여기에는 **이해당사자(관련 AI 행위자 포함)와의 정기적인 참여가 포함**된다.
	4.3 사고 및 오류는 영향을 받는 **커뮤니티를 포함하여 관련 AI 행위자에게 전달된다. 사고 및 오류를 추적하고, 이에 대응하며, 그로부터 복구하기 위한 프로세스를 준수하고 이를 문서화**한다.

실시간 CCTV 생체인식 AI 시스템과 관리 관련 검토

항목	검토의견
1.1 AI 시스템이 의도한 목적 및 **목표를 달성했는지 여부와 시스템의 개발 또는 배포를 진행해야 하는지 여부에 대한 결정**을 내린다.	목표 달성 여부에 대한 판단
1.2 문서화된 **AI 위험은 영향, 가능성, 가용 리소스 또는 방법에 따라 그 우선 순위가 지정**된다.	위험의 우선순위 지정
1.3 매핑 기능을 통해 식별된 **우선 순위가 높은 AI 위험에 대응하기 위한 방법을 개발, 계획 및 문서화**한다. 위험 대응 옵션에는 완화, 이전, 회피 또는 수용이 포함된다.	위험 관리 방법, 계획 문서화
1.4 AI 시스템의 **후속 취득자 및 최종 사용자 모두에 대한 부정적인 잔류 위험(완화되지 않은 모든 위험의 합계로 정의됨)을 문서화**한다.	후속 사용자 잔류 위험 문서화
2.1 **잠재적 영향의 규모 또는 가능성을 줄이기 위해 실행 가능한 비-AI 대체 시스템, 접근 방식 또는 방법과 함께 AI 위험을 관리하는 데 필요한 리소스를 고려**한다.	비 AI 대체 시스템 고려
2.3 **이전에 알려지지 않은 위험이 식별될 경우 해당 위험에 대응하고 그로부터 복구하기 위한 절차**를 준수한다.	새로운 우험 대응
2.4 의도한 목적과는 다른 성능 또는 결과를 나타내는 **AI 시스템을 대체, 해제 또는 비활성화하기 위한 메커니즘을 마련하고 관련 책무를 할당하고 파악**한다.	대체, 해체 비활성화 메커니즘
3.1 **제3자 리소스의 AI 위험 및 이점을 정기적으로 모니터**링하고 위험 제어를 적용하고 문서화한다.	제3자 리소스 위험 모니터링
3.2 **AI 시스템의 정기적 모니터링 및 유지 관리의 일환**으로 개발용으로 사용되는 사전 학습된 모델을 모니터링한다.	사전 학습 모델 모니터링
4.1 배포 후 AI 시스템에 대한 **모니터링 계획을 구현**한다. 여기에는 사용자 및 기타 **관련 AI 행위자의 의견을 수집하고 평가하기 위한 메커니즘, 이의 제기, 중단, 해제, 사고 대응, 복구 및 변경 관리가 포함**된다.	배포 후 행위자 의견 수렴, 모니터링
4.2 지속적인 개선 활동이 AI 시스템 업데이트에 통합되며, 여기에는 **이해당사자(관련 AI 행위자 포함)와의 정기적인 참여가 포함**된다.	이해당사자 포함 지속적 개선
4.3 사고 및 오류는 영향을 받는 **커뮤니티를 포함하여 관련 AI 행위자에게 전달된다. 사고 및 오류를 추적하고, 이에 대응하며, 그로부터 복구하기 위한 프로세스를 준수하고 이를 문서화**한다.	사고, 오휴 관련 행위자에게 전달, 추적 대응.

인공지능 영상검색·대상물 이동경로 추적 솔루션

- 과학기술정보통신부의 'AI 융합 국민안전 확보·신속대응 지원 공모사업' 일환으로 발굴됐다. AI 융합 기술을 활용해 기존 CCTV 관제 시스템에서 실종자 특징을 자동 분석하고 이동경로를 추적하는 것이 목표다. 시범 운영은 오는 17일부터 한 달간 진행된다. 올해 5월 사업 주관기관으로 선정된 마크애니는 제주자치도청과 제주경찰청, 민간기업 3곳과 연합체를 구성해 솔루션 개발을 이어왔다. 마크애니는 객체 분석 AI 모델 개발을 담당한다.
- **마크애니는 도내 방범용 CCTV로 수집한 학습 데이터 30만건을 가공해 1차 연도 목표인 AI 기반 객체 탐지·특징추출·비교 모델 구현을 10월 말 완료했다. 해당 모델은 현재 개발 중인 '실종자추적관리플랫폼(TOSS)'에 연동된다. 실종사건 발생 시 경찰이 플랫폼에서 실종자 사진과 인상착의, 수색 반경을 등록하면 AI는 유사도가 높은 대상자를 식별·추적한다.**
- 마크애니는 시범운영 기간 동안 **통신사 기지국 정보를 활용한 반경 지역 탐색 등 현장 밀착형 기능을 추가 보완할 계획**이다. 이번 운영 결과를 토대로 마크애니는 사업 2년 차 목표인 객체 탐지·추출·비교 모델을 결합한 AI 기반 객체 인식 최적화 모델을 개발할 계획이다.
- 최고 마크애니 대표는 "실종 사건은 골든타임 내 신속하게 발견하는 것이 중요하다"며 "이번 시범운영으로 AI 모델 성능을 검증하고 솔루션을 고도화해 사회 안전망 강화에 기여할 것"이라고 말했다. (김혜경 기자(hkmind9000@inews24.com)

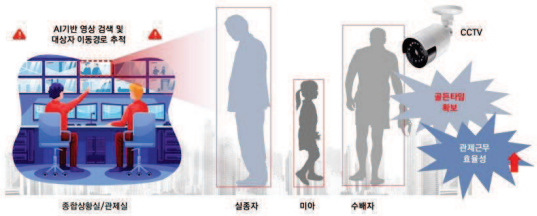
실종자추적관리플랫폼(TOSS) 로그인 화면.

- 마크애니는 도내 설치된 일반 방범용 CCTV로 수집한 30만 건 이상의 학습 데이터를 가공해 1차 연도 목표인 인공지능 기반 객체 탐지·특징추출·비교 모델 구현을 10월 말 완료했다. 해당 모델은 현재 개발 진행 중인 **실종자추적관리플랫폼(TOSS)**에 연동된다. **실종사건 발생 시 경찰이 플랫폼에서 실종자 사진과 의류 색상·장신구 등 인상착의, 수색 반경을 등록하면 인공지능은 유사도가 높은 대상자를 식별하고 추적**한다.

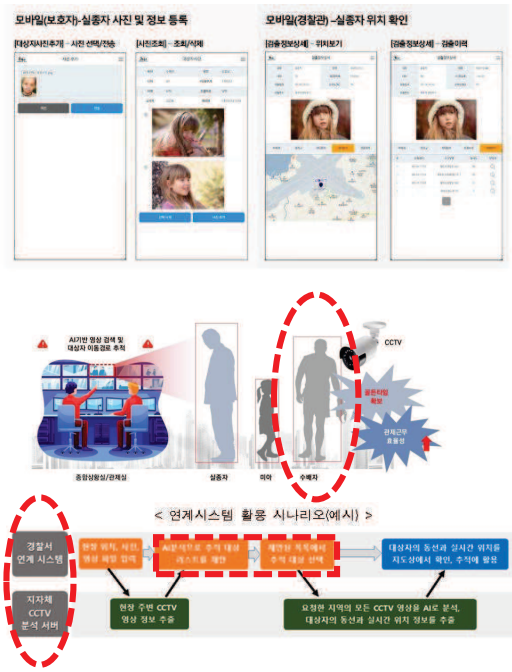
마크애니 사업 수주 보도(정보통신신문 2022.06.20)

- 크애니(대표 최고)는 과학기술정보통신부와 정보통신산업진흥원(NIPA)이 주관하는 'AI 융합 국민안전 확보 및 신속대응 지원 공모 사업'을 수주했다고 20일 밝혔다.
- 해당 사업은 미아, 치매노인 등 사회적 약자 실종사건 조기 대응을 위해 기존 방범용 CCTV 관제 시스템에서 신체적 특징과 소지품을 기반으로 실종자 수색 및 추적이 가능한 '인공지능(AI) 영상검색 및 대상물 이동경로 추적 솔루션' 개발을 목표로 한다. 개발 완료 후 제주도에 시범 적용해 솔루션을 고도화할 예정이다.
- 주관기관은 마크애니이며 제주특별자치도, 제주 경찰청, ㈜알체라, ㈜와이드큐브, ㈜스마트뱅크와 연합체를 구성해 기술 개발에 매진한다.
- **마크애니는 이번 사업으로 지자체 내 일반 방범용 CCTV 환경에서 객체 탐지, 위험 예측, 이동 경로 추적이 가능한 AI 기반 선별 관제 솔루션 개발로 실종사건 외 수배자 추적, 범죄위험 예측 등 각종 사건·사고를 선제 대응해 사회 안전망을 강화할 전망이다.**

AI 기반 선별 관제 솔루션 개발, CCTV 관제 시스템 진일보 기대(정보통신신문 2022.06.20)



AI 영상검색 및 대상물 이동경로 추적 솔루션 개념도. [자료=마크애니]



제기되는 법적 문제

- 대상자 조회는 개인위치추적/전통적인 출입, 조사 등과는 다른 법적 문제
  - 기존의 실종아동보호법은 개인위치추적이나, 출입, 조사 등이므로 기본권 침해의 문제가 크지 않음
  - 그러나, AI 시스템을 통한 실시간 식별, 추적은 과거와는 양상이 다름
- AI 활용 대상자 조회, 추적시
  - (추적대상 리스트 제안 → 제안된 목록에서 추적대상 선택)
  - 여러 명의 추천 대상자들에 대한 법적 보호는?
  - 오인 추적되는 경우의 법적 보호는?
- 경찰이나 지방자치단체의 오용 가능성에 대한 대응방안은?
  - 수사에 활용하려고 하는 경우
- AI 활용 실시간 생체인식 추적의 법적 근거가 마련되어 있는가?
  - 절차, 요건 등

합법성: 실시간 CCTV를 통한 실종자 수색이 허용되는가?

- 실종아동등의 보호 및 지원에 관한 법률(실종아동법)
  - 사전 지문 및 얼굴 등에 관한 정보 등록(제7조의 2), 실종아동 등의 지문등정보 등록(제7조의 3)
  - 정보연계시스템 구축, 운영(제8조)
    - 실종아동등의 신상정보를 작성, 취득, 저장, 송신·수신하는 데 이용할 수 있는 전문기관·경찰청·지방자치단체·보호시설 등과의 협력체계 및 정보네트워크
  - **실종아동등 신고·발견을 위한 정보시스템의 구축·운영(제8조의 2)**
  - 수색, 수사의 실시(제9조, 제9조의 2)
    - 실종아동등의 발생 신고를 접수하면 지체 없이 수색 또는 수사의 실시 여부를 결정
    - 즉시 현장에 출동시켜 주변 수색
    - 범죄로 인한 경우는 수사
    - 개인위치정보, IP 접속정보 제공 요청
    - 실종아동의 정보 문자, 인터넷, 방송으로 제공
  - 출입, 조사(제10조)
    - 관계인에 대하여 필요한 보고 또는 자료제출을 명하거나 소속 공무원으로 하여금 관계 장소에 출입하여 관계인이나 아동등에 대하여 필요한 조사 또는 질문
    - 출입·조사 또는 질문을 하려는 관계공무원은 그 권한을 표시하는 증표를 지니고 이를 관계인 등에게 내보여야

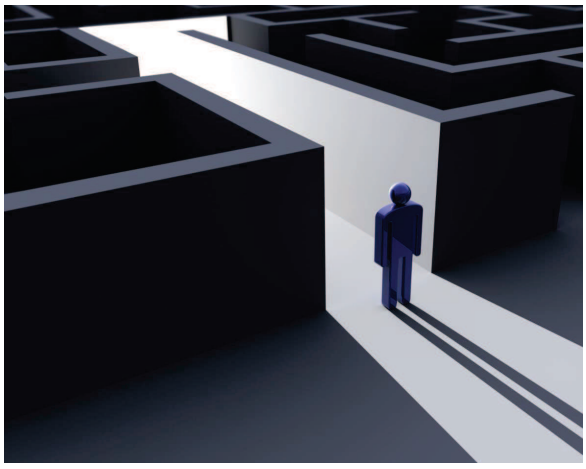
합법성: 실시간 CCTV를 통한 실종자 수색이 허용되는가?

- 불특정 다수인 이용시설에서 실종아동 등 조기 발견 지침과 조치(실종아동보호법 제9조의 3 제1항, **실종아동등 조기발견지침**)
  - **실종아동등 발생상황 전파와 경보발령**
    - **안내방송, 전광판 등**
  - **출입자 감시 및 수색실시**
    - 시설의 출입구에 종사자를 배치하여 출입자의 감시 등 필요한 조치를 취하되, 조치가 곤란하거나 불충분할 경우에는 즉시 관할 경찰관서의 장에게 신고하여야.
    - 관리주체는 이용자에게 공개된 장소뿐만 아니라 이용자의 접근이 제한되는 장소 및 시설에 대해서도 수색을 실시하여야

51

합법성에 중대한 의문

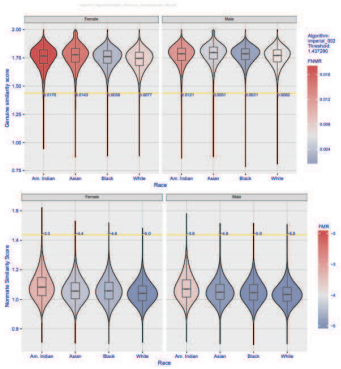
- 현행 실종아동보호법상 실종자 실시간 생체인식 추적은 허용된다고 보기 어려움
  - 현행 법은 개인위치정보 수집, 실종자 정보전파, 출입, 조사, 다중이용시설의 출입구 감시와 수색 등을 규정하고 있음.
  - 현행 규정상으로는 공개된 장소의 CCTV에서 불확실한 기법으로 실종자에 대한 AI 실시간 생체인식 추적기법을 사용하도록 허용한다고 보기는 어려움.
  - 현행 규정은 실종자 외의

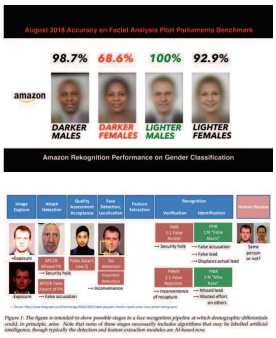


52

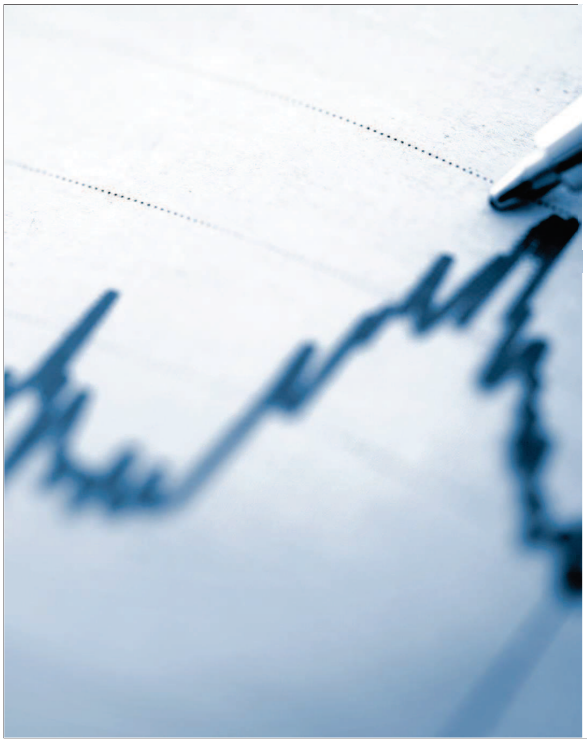
편향성

- 얼굴인식 기술의 편향성
- 학습 데이터
  - 액터(Actor : 학습데이터로 촬영된 주체)의 구성
    - 성별, 연령, 인종 등이 골고루 구성되었는지?
    - 장애인의 경우
- 추적 가능성의 측면에서
- 장애인의 경우 오인될 가능성





Face Recognition Vendor Test (FRVT), Part 3: Demographic Effects(NIST, 2019)



이해관계자의 참여

- 다양한 이해관계자 참여
- 오인된 피해자에 대한 구제절차
- 대중의 신뢰
- 추적 가능성에 대해 널리 알렸는지?
- 대중들이 쉽게 파악할 수 있게 되어 있는지?
- 제3자의 감시, 감독이 가능한지?
- 비용 대비 효율성이 있는지?
- 시민들과 협의가 되었는지?

기본권 영향 평가 등

- 기본권 영향
  - 영향을 받을 수 있는 기본권
- 오용에 대한 대비
  - 수사 목적 활용, 기타 오용 가능성에 대한 대응

한국통신기술협회(TTA) 인공지능 신뢰성 인증

인공지능 시스템 신뢰성 제고를 위한 요구사항 목록

요구사항	내용
REQ.01	인공지능 시스템에 대한 위험관리 계획 및 수행 - 인공지능 시스템 생명주기에 걸쳐 나타날 수 있는 위험 요소 분석 - 위험 요소를 제거 및 방지하거나 영향을 완화하기 위한 방안 마련
REQ.02	인공지능 체계 구성 - 인공지능 에 대한 지침 및 규정을 수립 - 인공지능 를 위한 조직을 구성하고 인력 구성 검토 - 인공지능 체계가 수립한 대로 이행되고 있는지 감독 - 인공지능 조직이 신규 및 기존 시스템의 차이점 분석
REQ.03	인공지능 시스템의 신뢰성 테스트 계획 수립 - 인공지능 시스템의 특성을 고려한 테스트 환경 설계 - 인공지능 시스템의 테스트 설계에 필요한 협의 체계 구성
REQ.04	인공지능 시스템의 추적가능성 및 변경이력 확보 - 인공지능 시스템의 의사결정에 대한 추적 방안 수립 - 데이터의 변경 이력을 확보하고, 데이터 변경이 미치는 영향 관리
REQ.05	데이터의 활용을 위한 상세 정보 제공 - 데이터의 명확한 이해와 활용을 지원하는 메타데이터 등의 상세 정보 제공 - 데이터의 출처 기록 및 관리
REQ.06	데이터 견고성 확보를 위한 이상 데이터 점검 - 이상 데이터의 식별 및 정상 여부 점검 - 데이터 공격에 대한 방어 수단 준비
REQ.07	수집 및 가공된 학습 데이터의 편향 제거 - 데이터 수집 시, 인적·물리적 요인으로 인한 편향 완화 방안 마련 - 학습에 사용되는 특성(feature)을 분석하고 선정 기준 마련 - 데이터 라벨링 시, 발생 가능한 인적 편향을 확인하고 방지 - 데이터의 편향 방지를 위한 샘플링 수행

인공지능 시스템 신뢰성 제고를 위한 요구사항 목록

요구사항	요구사항 명
REQ.08	인공지능 오픈소스 라이브러리의 보안성 및 호환성 점검 - 인공지능 오픈소스 라이브러리 업데이트 수행 여부 확인 - 인공지능 오픈소스 라이브러리 위험요소 점검 후 안전성 확인
REQ.09	인공지능 모델의 편향 제거 - 모델 편향을 제거하는 기법 적용을 학습 전, 학습 과정 중, 학습 이후로 나누어 확인
REQ.10	인공지능 모델 공격에 대한 방어 대책 수립 - 모델 추출 공격에 대한 방어 방안 수립 - 모델 회피 공격에 대한 방어 방안 수립
REQ.11	인공지능 모델 명세 및 추론 결과에 대한 설명 제공 - 모델의 명세를 투명하게 제공하기 위해, 인공지능 모델 상세 문서 제공 - 사용자가 모델 추론 결과의 도출 과정을 수용할 수 있는 근거 제공 - 인공지능 모델 추론 결과에 대한 설명 제공 필요성 확인
REQ.12	인공지능 시스템 구현 시 발생 가능한 편향 제거 - 소스 코드 및 오출력, 사용자 인터페이스로 인한 편향 제거 기법 적용 확인
REQ.13	인공지능 시스템의 안전 모드 구현 및 문제발생 알림 절차 수립 - 공격, 성능 저하 및 사회적 이슈 등의 문제발생 시 대응 가능한 안전 모드 적용 - 인공지능 시스템에서 문제발생 시, 시스템은 해당 문제를 운영자에게 전달하는 기능 수행
REQ.14	인공지능 시스템의 설명에 대한 사용자의 이해도 제고 - 인공지능 시스템 사용자의 특성 및 제약사항 분석 - 사용자 특성에 따른 충분한 설명 제공
REQ.15	서비스 제공 범위 및 상호작용 대상에 대한 설명 제공 - 인공지능 서비스의 올바른 사용을 유도하기 위한 설명 제공 - 사용자와의 상호작용 대상을 명시하는 설명 제공

요구사항 01 인공지능 시스템에 대한 위험관리 계획 및 수행

01-1 인공지능 시스템 생명주기에 걸쳐 나타날 수 있는 위험 요소를 분석하였는가?	• 01-1a 인공지능 시스템의 위험 요소를 도출하고 이의 파급효과를 파악하였는가?
01-2 위험 요소를 제거 및 방지하거나 영향을 완화하기 위한 방안을 마련하였는가?	• 01-2a 위험 요소 제거 방안을 도출하고 파급효과가 감소하였는지 확인하였는가?

요구사항 02 인공지능 governance 체계 구성

- 02-1 인공지능 에 대한 지침 및 규정을 수립하였는가?
  - 02-1a 내부적으로 준수해야 할 인공지능 에 대한 지침 및 규정을 마련하였는가?
- 02-2 인공지능 를 위한 조직을 구성하고 인력 구성에 대해 검토하였는가?
  - 02-2a 인공지능 를 위한 조직을 구성하였는가?
  - 02-2b 인공지능 를 위한 조직은 충분히 훈련된 인력으로 구성하였는가?
- 02-3 인공지능 체계가 올바로 이행되고 있는지 감독하고 있는가?
  - 02-3a 인공지능 에 대한 내부 지침 및 규정 준수 여부를 감독하고 있는가?
- 02-4 인공지능 조직이 신규 및 기존 시스템의 차이점을 분석하였는가?
  - 02-4a 이용 빈도가 낮은 타 시스템의 개선 및 통폐합을 통해 구현 가능한지 분석하였는가?

요구사항 03 인공지능 시스템의 신뢰성 테스트 계획 수립

- 03-1 인공지능 시스템의 특성을 고려한 테스트 환경을 설계하였는가?
  - 03-1a 테스트 환경 결정 시 인공지능 시스템의 운영환경을 고려하였는가?
  - 03-1b 가상테스트 환경이 필요한 인공지능 시스템의 경우, 시뮬레이터를 확보하였는가?
- 03-2 인공지능 시스템의 테스트 설계에 필요한 협의 체계를 구성하였는가?
  - 03-2a 인공지능 시스템의 기대 출력을 결정하기 위한 협의 체계를 구성하였는가?
  - 03-2b 설명가능성 및 해석가능성 확인을 위한 사용자 평가단을 구성하였는가?

• 2 데이터 수집 및 처리
• 요구사항 04 데이터의 활용을 위한 상세 정보 제공

- 04-1 데이터의 명확한 이해와 활용을 지원하는 상세한 정보를 제공하는가?
  - 04-1a 정제 전과 후의 데이터 특성을 설명하였는가?
  - 04-1b 학습 데이터와 메타데이터metadata를 구분하고 각 명세자료를 확보하였는가?
  - 04-1c 보호변수protective attribute의 선정 이유 및 반영 여부를 설명하였는가?
  - 04-1d 라벨링 작업자를 위해 교육을 시행하고 작업 가이드 문서를 마련하였는가?
- 04-2 데이터의 출처는 기록 및 관리되고 있는가?
  - 04-2a 신뢰할 수 있는 출처로부터 제공되는 데이터셋을 사용하였는가?
  - 04-2b 오픈소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?

CCTV 연계 실시간 생체인식 AI 시스템 - ALTAI에 의한 평가 항목

- 인간의 주체성, 자율성, 인간의 감독
- 기술적 견고성 및 안전성
  - 복원력
  - 정확성
  - 신뢰성, 대체 계획
  - 재현성
- 개인정보보호 및 데이터 거버넌스
- 투명성
  - 추적성
  - 설명가능성
  - 의사소통
- 다양성, 차별금지, 공정성
  - 불공정한 편견방지
  - 접근성과 유니버설 디자인
  - 이해관계자 참여
- 사회 및 환경복지
  - 사회 전반이나 민주주의에 미치는 영향
- 책임
  - 감사가능성

CCTV 연계 실시간 생체인식 AI 시스템 - 기본권 영향 평가

- 충돌하는 기본권
  - 실종자, 가족의 기본권 vs 촬영되는 사람의 기본권
  - 촬영되는 개인, 집단
- AI 시스템 오남용시 충돌하는 기본권
  - 수사기법으로 활용될 가능성
  - 목적 외 유출, 오남용될 가능성
  - 표현의 자유 제한 가능성
  - 사생활의 비밀 제한 가능성

CCTV 연계 실시간 생체인식 AI 시스템

투명성, 설명가능성

- 문서화
- 공개
  - 운영지침
  - 세부기준
  - 운영 현황
  - 사고와 위험
- 설명

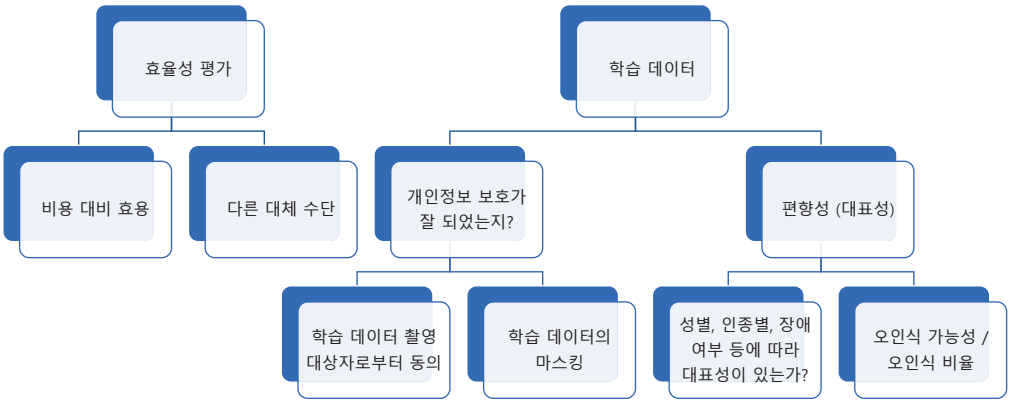
책임성과 오용에 대한 대비

- 오용과 관련한 책임성
- 내부 감독과 외부의 독립적인 감사
- 오용에 대한 이의 신청, 권리 구제 등

경쟁 가능성

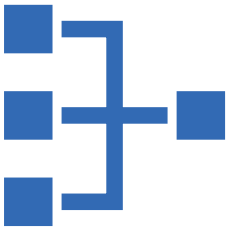
- 경쟁 업체와의 형평성
- 연구 결과와 지적재산권

CCTV 연계 실시간 생체인식 AI 시스템



CCTV 연계 실시간 생체인식 AI 시스템

- 학습데이터의 액터(Actor) 구성은 적절하게 이루어졌는지?
  - 알고리즘의 민감성과 연관이 있을 수 있음.
- 알고리즘에 유사한 객체, 사람으로 인식되는 객체나 사람은 추적이 이루어질 것이므로, 유사성이 과도하게 측정되면 인권 침해 가능성은 높아짐.
  - 행동의 특징이 있는 장애의 경우 유사성이 과도하게 측정될 가능성이 높다면 과도하게 추적되고, 인권 침해될 가능성은 높아짐.
  - 특징적인 용모나 객체를 지닌 사람들(긴머리, 삭발, 특정한 복장 등)이 과도하게 추적될 가능성
- 정확도를 낮출 경우 false positive의 가능성이 높아짐. 오인되어 추적될 가능성이 높아짐. 반면 정확도를 높일 경우는 false negative의 가능성이 높아져서 효율성이 떨어질 수 있음. 결국, 이 시스템이 공공 장소에 설치된 CCTV 전부에 대해서 활용될 수 있게 되면 사생활 침해의 위험, 표현의 자유에 대한 위험 등이 커질 수 있음.



그 밖에 남는 문제들...

- 다양한 이해관계자 참여가 이루어졌는지?
- 오인된 피해자에 대한 구제절차가 있는지?
- 대중의 신뢰를 얻을 수 있는지?
- CCTV를 통한 추적 가능성에 대해 널리 알렸는지?
- 보행하는 시민들이 쉽게 파악할 수 있게 되어 있는지?
- 오용의 위험에 대한 대비는 충분한지?
- 제3자의 감시, 감독이 가능한지?

AI 시스템에 대한 표준화 논의에서 고려할 점

사회기술적 요소가 강하다는 점을 고려해야 함

- 기술적인 측면 뿐만 아니라, 사회환경에 미치는 영향이 큼.
- 불투명성, 데이터 특성 등을 통해 알고리즘의 확대, 심화되는 특성 등으로 사회적 영향 고려가 매우 중요

소비자, 기본권 영향 등에 대한 충분한 고려

- 표준 작성 과정에서부터 이들의 적극적인 참여 필요함.
- 이들 분야의 전문가들이 참여하도록 해야 함.

국제적인 표준과 조화

- 미국, EU는 표준의 국제화를 추진함. 미국,EU는 사회기술적 요소를 중요하게 여기고, 이를 중요 사항으로 포함시키고 있음.
- 우리나라 표준에서 이런 부분을 도외시할 경우, 국제적으로 고립될 가능성이 높음.

【 토론 1 】

송경호

(연세대 정치학과 BK21 박사후연구원)

【 토론 2 】

이현경

(KISDI 지능정보사회정책연구실, 부연구위원)

【 토론 3 】

박소영

(국회입법조사처 입법조사관, 변호사)