"AI & Security Guidelines"

- 인공지능과 보안가이드라인 V1.0 -

(In The Artificial Intelligence Society)

2023.07.15

Sangshik, Min
(mikado22001@yahoo.co.kr)

FPRI
Future Policy Research Institute

개요

"**AI 윤리원칙**" 수립 → **"AI윤리점검체크(리스트)"**
→ "**AI위험관리**"계획 및 세부기획/개발에 대한 "**AI영향평가**" 필요

* AI는 PDCA 관점에서 관리되어야 함

□ AI기반 사회

"**AI 윤리원칙**" 수립 → "**AI윤리점검체크(리스트)**"
→ "**AI위험관리**"계획 및 세부기획/개발에 대한 "**AI영향평가**" 필요

- AI에 기반한 자동화된 결정으로 인간스스로의 통제력이 감소된 사회에서 살아가야 할것
  : AI는 인간을 보조하는 단순 범용기술이 아닌,
  생산성, 프라이버시, 평등, 차별, 노동시장, 감시 등 많은 사회적이슈와 연결되어있음

- **AI윤리원칙**의 수립 후 현실에의 적용을 위해서는, 실행가능한 제도/절차를 마련해야 함
  : **AI윤리준수 점검리스트**나 AI관련 윤리/법제도 **영향평가** 수행 등

과학과 신기술에 대한 윤리 그룹의 윤리적 원칙들		
인간의 존엄성	정의, 공정, 연대	보안, 안전, 심신의 건전성
자율성	민주주의	데이터 보호와 프라이버시
책임	법치와 책무성	지속가능성

출처: European Group on Ethics in Science and New Technologies (2018) "Statement on Artificial Intelligence, Robotics and Autonomous Systems".

AI Control

□ AI & Security (요약)

- 데이터와 프로그램에 대한 통제프로세스 필수

거버넌스/수집/제공/개발/모니터링 等

[통제 일반]
- 프로그램의 통제를 위한 테스트 로직 강화 필수

[AI에 대한 보안통제] – 가이드라인 필요
- 거버넌스 검토/운영
- 데이터 검증 (개인정보)
- 프로그램 검증 (SW 취약점 검증)

// TODO

[별첨] FTC launches investigation into ChatGPT (2023년 7월 13일) - 미국 연방거래위원회(FTC) ChatGPT의 법위반 조사
[별첨] Inside the White-Hot Center of A.I. Doomerism (2023년 7월 11일) – AI 업체 Anthropic
[별첨] NIST AI위험관리 프레임워크 (2023년 1월 26일)
[별첨] 유네스코 인공지능 윤리권고 (2021년 11월 25일)

[별첨] FTC launches investigation into ChatGPT (2023년 7월 13일)

*** 미국 연방거래위원회(FTC) ChatGPT의 법위반 조사**

https://www.youtube.com/watch?v=SNm223KY78g

미국의 공정거래위원회 격인 연방거래위원회(FTC)가 생성형 인공지능(AI) 챗GPT를 개발한 오픈AI에 대해 소비자보호법 위반 여부 조사에 착수

- FTC가 이번 주 오픈AI 측에 보낸 20쪽 분량의 공문
  개인정보 보호에 기만적인 행위가 있었는지, 소비자에게 해로운 관행이 있는지 등을 조사하기 위해
  (훈련 데이터)
  △ 오픈AI가 챗GPT를 교육하는 데 사용한 자료
  △ 해당 자료의 출처와 취득 방식
  (출력 데이터)
  △ 올 3월 오픈AI가 공지한 사용자 개인정보 유출 사고 관련 자료 등
  △ 챗GPT가 실존 인물에 관한 거짓 정보를 제공해 회사에 불만이 접수된 사례
  (통제)
  △ 이에 대한 회사 측의 조치

- FTC는 오픈AI가 소비자보호법을 위반했다고 판단할 경우 벌금을 부과하거나 시정을 명령가능

- 샘 올트먼 오픈AI의 공동창업자 겸 최고경영자(CEO·사진)는 올 5월 미국 의회에서 처음으로 열린 AI 청문회에 참석해 "**AI가 내년 대선에서 거짓 정보를 퍼뜨릴 수 있다**"며 "AI를 규제해야 한다"고 촉구한 바 있음, 이를 두고 정보기술(IT) 업계에선 AI 규제가 불가피한 상황에서 이미 경쟁에 앞선 오픈AI가 규제 논의를 앞당겨 시장을 선점하려는 전략을 택한 것이라는 지적

- 올트먼 CEO는 이날 자신의 트위터를 통해 "우리의 기술이 고객 친화적이고 안전하다는 것은 우리에게도 매우 중요하다"며 "우리는 관련 법을 준수하고 있다고 확신하며, FTC의 조사에 협조할 것"이라고 밝혔다.

The New York Times

[별첨] Inside the White-Hot Center of A.I. Doomerism (2023년 7월 11일)

Anthropic, a safety-focused A.I. start-up, is trying to compete with ChatGPT while preventing an A.I. apocalypse. It's been a little stressful.

Constitutional AI

주의 깊게 제어할 필요

헌법적 AI는 AI 모델에 문서화된 원칙 목록(헌법)을 제공하고 가능한 한 이러한 원칙을 따르도록 지시

인공 지능 → 일반 인공 지능(AGI) →

시스템이 우리를 장악하고 파괴할 수 있음
(2028~2033년 쯤)

①**Constitutional AI를 사용**하여 높은 수준에서 언어를 이해하도록 훈련시키면 시스템이 자체 규칙을 위반하는 시기를 알거나 덜 강력한 모델이 허용했을 수 있는 잠재적으로 유해한 요청을 종료할 수 있음

②**사고 실험**의 형태로 강력한 AI 시스템을 에 대한 도덕적 사례

https://www.nytimes.com/2023/07/11/technology/anthropic-ai-claude-chatbot.html

[별첨] NIST AI위험관리 프레임워크 (2023년 1월 26일)

美 NIST「AI 위험관리 프레임워크(AI RMF) 1.0」분석 및 시사점 보고서 - 한국지능정보사회진흥원(NIA) 참고

AI 위험관리 프레임워크 1.0(이하 AI RMF)

- 미 의회 : '국가인공지능 이니셔티브법(National Artificial Intelligence Initiative Act of 2020 (P.L. 116-283)) '20년 제정, 이에 따른 RMF 수립
- RMF 목적 : 모든 분야・규모의 기업・조직이 AI 위험을 해결할 수 있도록 유연하고 체계적이며 측정 가능한 프로세스를 제공
  /협업과 참여/보편적 활용/조직의 전체적 위험관리에 포함/결과 중심/기존체제 내 적용/빠른업데이트/
- 핵심내용 : 신뢰할 수 있는 AI 시스템의 7가지 특성(안전성, 책임·투명성 등)을 제시하고,
  이를 위한 조직의 핵심 기능을 4가지 영역(거버넌스, 매핑(위험 식별), 측정, 관리)에서 제시

OECD의 '인공지능 시스템 분류 및 프레임워크'(2022)

- **인간과 지구** | 사용자 및 이해관계자 입장에서 인공지능이 미치는 영향을 분석 할 수 있음
- **경제적 맥락** | 산업 분야, 비즈니스 모델, 성숙도 등을 분석
  ※ NIST는 AI 시스템 프로세스에 따라 분류해 경제적 맥락을 제외하고 응용 프로그램을 추가
- **데이터 및 입력** | 학습 데이터의 출처, 구조, 형식 등을 파악할 수 있으며, 입력 데이터의 동적 특성(수시, 실시간 업데이트 등)의 분석도 가능
- **모델** | 모델의 특징(규칙 기반, 머신러닝 등), 구축 방법뿐만 아니라 모델이 추론하는 방식 등 모델 전반에 대한 이해가 가능
- **작업(task)과 출력** | 시스템이 수행하는 서비스(인지, 예측 등)의 종류를 분석하고, 적용 분야(컴퓨터 비전, 언어 기술)등에 대한 분석이 가능

출처 : 우상근(2022.4), OECD 인공지능 시스템 분류 프레임워크 분석 및 시사점, AI REPORT 2022-1

OECD Aritificial Intelligence in Society(2019)

OECD Recommendation on AI(2019) ; ISO/IEC 22989(2022)

ISO/IEC TR 24368:2022

[AI시스템의 위험]

학습 데이터

- 사회적 기술 : 기존 소프트웨어와 비교했을 때 AI는 시간이 지남에 따라 **예상치 못한 데이터로 훈련**되어 시스템에 영향을 미칠 수 있는 '사회적 기술(Socio-tech)'로 사회 역학과 인간의 행동에 의해 영향받기 쉬움
- 데이터 품질 문제 : AI 시스템을 구축하는 데 사용되는 데이터가 AI 시스템의 상황 또는 의도된 사용을 정확하게 대표하지 못할 수 있으며 실제 상황을 반영하지 않거나 사용 가능하지 않을 수 있으며 유해한 편견 및 기타 **데이터 품질 문제**가 AI 시스템의 신뢰성에 영향을 미칠 수 있음
- 데이터 의존성 : AI 시스템은 **학습데이터에 대한 종속성, 의존도**가 높으며 데이터 학습 중 의도하거나 의도치 않은 변경으로 인해 AI 시스템의 성능이 근본적으로 달라질 수 있음
- 불확실성 : 연구를 발전시키고 성능을 개선할 수 있는 **사전 학습 모델을 사용**할 경우 통계적 불확실성이 높아지고 편향 관리, 과학적 타당성 및 재현성에 문제가 발생할 수 있으며 대규모 사전 훈련 모델의 경우 새로운 속성에 대한 오류를 예측하는 것이 더욱 어려움

알고리즘

- 규모와 복잡성 : AI 시스템은 **수십억 개 또는 수조 개의 결정 지점**을 포함하고 있으며 이는 일반적인 소프트웨어 응용 프로그램과 함께 작동하여 더욱 복잡
- 제어 및 유지 보수의 어려움 : AI 시스템은 전통적인 코드 개발과는 다른 제어를 받기 때문에 정기적인 AI 기반 소프트웨어 테스트를 수행하거나 무엇을 **테스트해야 하는지 결정하기 어려울** 수 있으며 데이터, 모델 또는 개념의 변화로 인해 빈번한 유지 보수 관리가 필요할 수 있음
- 문서화 관리의 어려움 : 소프트웨어의 테스트 표준이 개발되지 않아 가장 단순한 경우를 제외한 모든 경우에 대해 기존의 소프트웨어에서 예상되는 표준에 대한 **AI 기반 수행 기준을 문서화하기 어려움**

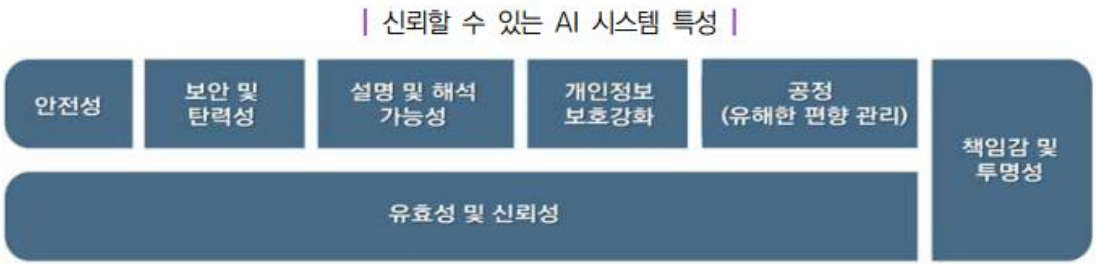
학습데이터 전처리 중요

- 개인정보 위험 : 대량의 데이터를 처리하고 분석하는 과정에서 민감한 개인 정보를 포함할 가능성이 높아 **개인정보보호 위험이 발생** 가능

[AI시스템 생명주기]


응용 프로그램
계획및설계
감사 및 영향 평가 포함
데이터 및 입력
데이터수집 및처리
내부 및 외부 검증 포함
AI 모델
모델구축 및사용
모델테스트 포함
AI 모델
확인및검증
모델테스트 포함
작업 및 출력
배포및사용
통합, 규정 준수 테스트 및 검증 포함
인간과 지구
사용및영향
감사 및 영향 평가 포함
응용 프로그램
운영및 모니터링
감사 및 영향 평가 포함

[신뢰할 수 있는 AI시스템]



[신뢰할 수 있는 AI시스템 개발을 위한 필요기능]

[신뢰할 수 있는 AI시스템 개발을 위한 필요기능] - 거버넌스

범주(기능)	하위범주
거버넌스 1 AI 위험 매핑, 측정 및 관리와 관련된 조직 전반의 정책, 프로세스, 절차 및 수행 기준이 투명하고 효과적으로 구현된다.	**1.1** AI와 관련된 법적 및 규제적 요구 사항을 이해, 관리 및 문서화한다.
	1.2 신뢰할 수 있는 AI의 특성을 조직적 정책, 프로세스, 절차 및 수행 기준에 통합한다.
	1.3 조직의 위험 허용 범위를 기반으로 위험관리 활동의 수준을 결정하기 위한 프로세스, 절차 및 수행 기준을 마련한다.
	1.4 위험관리 프로세스 및 그 결과는 투명한 정책, 절차 및 조직의 위험 우선 순위를 기반으로 한 기타 제어를 통해 설정된다.
	1.5 위험관리 프로세스 및 그 결과에 대한 지속적인 모니터링 및 정기적인 검토를 계획하고 주기적 검토 빈도를 포함하여 조직의 역할 및 책무를 명확하게 정의한다.
	1.6 AI 시스템 인벤토리를 작성하는 메커니즘을 구축하며 조직의 위험 우선 순위에 따라 리소스를 할당한다.
	1.7 위험성을 높이거나 조직의 신뢰성을 떨어뜨리지 않는 방식으로 AI 시스템을 안전하게 해제하고 단계적으로 중단하기 위한 프로세스 및 절차를 마련한다.
거버넌스 2 적절한 부서 및 직원에게 AI 위험을 매핑, 측정 및 관리하기 위한 권한을 부여하고 교육을 받을 수 있도록 하는 책임 구조를 구축한다.	**2.1** AI 위험을 매핑, 측정 및 관리하는 것과 관련된 역할, 책임 및 커뮤니케이션 내용을 문서화하고 이를 조직 전반의 부서 및 개인에게 명확히 인지시킨다.
	2.2 조직 내 직원 및 파트너는 관련 정책, 절차 및 계약에 따라 그들의 의무와 책임을 수행할 수 있도록 AI 위험관리 교육을 받는다.
	2.3 조직의 경영진은 AI 시스템의 개발 및 배포와 관련된 위험에 대해 의사 결정할 책임을 가진다.
거버넌스 3 주기 전반에 걸쳐 AI 위험을 매핑, 측정 및 관리하기 위해 인력 다양성, 형평성, 포용성 및 접근성 프로세스의 우선 순위를 설정한다.	**3.1** 주기 전반에 걸쳐 AI 위험을 매핑, 측정 및 관리하는 의사 결정을 내릴 때 다양한 부서로부터 정보를 얻는다.
	3.2 인간-AI 구성 및 AI 시스템 감독과 관련된 역할과 책임을 정의하고 차별화하기 위한 정책 및 절차를 마련한다.

범주(기능)	하위범주
거버넌스 4 조직 내 부서는 AI 위험을 고려하고 해당 내용에 대해 커뮤니케이션하는 문화를 구축하기 위해 노력한다.	**4.1** 잠재적인 악영향을 최소화하기 위해 AI 시스템을 설계, 개발, 배포 및 사용하는 데에 있어 비판적 사고 및 안전 우선 주의 방식을 장려하기 위한 조직적 정책 및 수행 기준을 마련한다.
	4.2 조직 내 부서는 설계, 개발, 배포, 평가 및 사용하는 AI 기술의 위험 및 잠재적 영향을 문서화하고 그 영향에 대해 보다 광범위하게 커뮤니케이션한다.
	4.3 AI 테스트, 사고 식별 및 정보 공유를 위한 조직적 수행 기준을 마련한다.
거버넌스 5 AI 행위자의 강력한 참여를 유도하기 위한 프로세스를 마련한다.	**5.1** AI 위험과 관련된 잠재적인 개인적/사회적 영향에 대해 AI 시스템을 개발 또는 배포한 부서 외부의 피드백을 수집, 고려, 우선 순위 지정 및 통합하기 위한 조직적 정책 및 수행 기준을 마련한다.
	5.2 AI 시스템을 개발 또는 배포한 부서가 관련 AI 행위자의 피드백을 시스템 설계 및 구현에 정기적으로 통합할 수 있도록 하는 메커니즘을 구축한다.
거버넌스 6 제3자의 소프트웨어, 데이터 및 기타 공급망 문제로 발생하는 AI 위험 및 이점을 해결하기 위한 정책 및 절차를 마련한다.	**6.1** 제3자의 지적재산권 또는 기타 권리 침해를 포함하여 제3자 기관과 관련된 AI 위험을 해결하기 위한 정책 및 절차를 마련한다.
	6.2 위험성이 높은 것으로 간주되는 제3자의 데이터 또는 AI 시스템의 고장 및 사고를 해결하기 위한 비상 프로세스를 마련한다.

[신뢰할 수 있는 AI시스템 개발을 위한 필요기능] - 매핑

범주(기능)	하위범주
매핑 1 상황을 설정 및 파악한다.	**1.1** 목적, 잠재적으로 유익한 용도, 상황 별 법률, 규범, 기대치, AI 시스템이 배포될 예상 조건을 파악하고 이하를 고려해 문서화한다. 〈고려 사항〉 · 사용자의 특정 집합 또는 유형(기대치 포함) · 시스템이 개인, 커뮤니티, 조직, 사회 및 지구에 미치는 잠재적인 긍정적/부정적 영향 · 개발 또는 제품 AI 주기 전반에 걸친 AI 시스템의 목적, 용도 및 위험에 관한 가정 및 관련 제한 사항 · 관련 TEVV 및 시스템 지표
	1.2 학제 간 AI 행위자, 기능, 기술 및 상황 설정을 위한 역량은 인구통계학적 다양성, 광범위한 도메인 및 사용자의 전문 지식을 반영하며, 이들의 참여는 문서화된다. 학제 간 협업 기회는 우선순위로 지정된다.
	1.3 AI 기술에 대한 조직의 사명 및 목표를 파악하고 문서화한다.
	1.4 비즈니스 가치 또는 비즈니스 상황을 명확하게 정의하거나 (기존의 AI 시스템을 평가하는 경우) 재평가한다.
	1.5 조직의 위험 허용 범위를 파악하고 문서화한다.
	1.6 관련 AI 행위자로부터 시스템 요구 사항(예: 사용자의 개인정보를 보호해야 하는 시스템)을 도출하고 파악한다. 설계 의사 결정 시 AI 위험을 해결하기 위해 사회-기술적 영향을 고려한다.
매핑 2 AI 시스템을 분류한다	**2.1** AI 시스템이 지원하는 작업을 구현하는 데 사용되는 특정 작업 및 방법을 정의한다(예: 분류자, 생성 모델, 추천자)
	2.2 AI 시스템의 정보 한계 및 인간이 시스템 결과를 활용하고 감독하는 방법에 관한 정보가 문서화된다. 문서화를 통해 관련 AI 행위자가 의사 결정을 내리고 후속 조치를 취하기 위한 충분한 정보를 제공할 수 있다.
	2.3 실험 설계, 데이터 수집 및 선택(예: 가용성, 대표성, 적합성), 시스템 신뢰성 및 구성 타당성과 관련된 항목을 포함하여 과학적 무결성 및 TEVV 고려 사항을 식별하고 문서화한다
매핑 3 적절한 벤치마크와 비교하여 AI 기능, 대상 용도, 목표, 예상 이점 및 비용을 파악한다.	**3.1** AI 시스템의 기능 및 성능에 대한 잠재적 이점을 조사하고 문서화한다.
	3.2 잠재적/실제적 AI 오류 또는 시스템의 기능 및 신뢰성(조직의 위험 허용 범위와 연관됨)으로 인해 발생하는 비금전적 비용을 포함한 잠재적 비용을 조사하고 문서화한다.
	3.3 대상 응용 프로그램 범위는 시스템 기능, 설정된 상황 및 AI 시스템 분류를 기반으로 지정 및 문서화된다.
	3.4 AI 시스템 성능, 신뢰성, 관련 기술 표준 및 인증에 대해 운영자 및 실무자를 숙련시키는 프로세스를 정의, 평가 및 문서화한다.
	3.5 거버넌스 기능의 조직적 정책에 따라 감독 프로세스를 정의, 평가 및 문서화한다.
매핑 4 제3자의 소프트웨어 및 데이터를 포함하여 AI 시스템의 모든 구성 요소에 대한 위험 및 이점을 매핑한다	**4.1** 제3자의 지적재산권 또는 기타 권리 침해 위험과 마찬가지로 AI 기술과 구성 요소의 법적 위험(제3자의 데이터 또는 소프트웨어 사용 포함)을 매핑하는 방법을 확립하고 이를 준수하여 문서화한다.
	4.2 제3자의 AI 기술을 포함하여 AI 시스템 구성 요소에 대한 내부 위험 통제를 식별하고 문서화한다.
매핑 5 개인, 그룹, 커뮤니티, 조직 및 사회에 대한 영향을 특성화 한다	**5.1** 예상 용도, AI 시스템의 과거 용도, 공개 사건 보고, AI 시스템을 개발 또는 배포한 팀에 대한 외부 피드백 또는 기타 데이터를 기반으로 식별된 각 영향(잠재적으로 긍정적인 또는 부정적인 영향 모두)에 대한 가능성과 규모를 식별하고 문서화한다.
	5.2 관련 AI 행위자의 정기적인 참여를 지원하고 긍정적/부정적/예상치 못한 영향에 관한 피드백을 통합하기 위한 절차 및 인력을 구축하고 이를 문서화한다.

[신뢰할 수 있는 AI시스템 개발을 위한 필요기능] - 측정

범주(기능)	하위범주
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 지표를 선택한다. 측정하지 않거나 측정할 수 없는 위험 또는 신뢰도 특성을 적절히 문서화한다.
	1.2 오류 보고서 및 커뮤니티에 대한 잠재적 영향을 포함하여 AI 지표의 적절성 및 기존 제어의 효율성을 정기적으로 평가 및 업데이트한다.
	1.3 시스템의 일선 개발자 또는 독립 평가자의 역할을 하지 않은 내부 전문가를 정기적 평가 및 업데이트에 참여시킨다. 도메인 전문가, 사용자, AI 시스템을 개발 또는 배포한 팀의 외부 AI 행위자 및 영향을 받는 커뮤니티는 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 문서화한다.
	2.2 인간 피실험자와 관련된 평가는 관련 요구 사항(인간 피실험자 보호 포함)을 충족하고 모집단을 대표한다.
	2.3 AI 시스템의 성능 또는 보증 기준을 정성적 또는 정량적으로 측정하고 배포 조건과 유사한 조건에서 입증한다. 조치를 문서화한다.
	2.4 매핑 기능에서 식별된 AI 시스템 및 구성 요소의 기능과 동작은 제조 시 모니터링된다.
	2.5 배포할 AI 시스템이 타당하고 신뢰할 수 있는지를 입증한다. 기술 개발 조건 이외의 일반화 한계를 문서화한다.
	2.6 매핑 기능에서 식별되는 안전 위험에 대해 AI 시스템을 정기적으로 평가한다. 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 반영한다.
	2.7 매핑 기능에서 식별된 AI 시스템의 보안 및 탄력성을 평가 및 문서화한다.
	2.8 매핑 기능에서 식별된 투명성 및 책임과 관련된 위험을 조사하고 문서화한다.
	2.9 AI 모델을 설명, 검증 및 문서화해야 하며 책임 있는 사용과 거버넌스 기능에 대해 알리기 위해 AI 시스템 결과를 매핑 기능을 통해 식별한 상황 내에서 해석해야 한다.
	2.10 매핑 기능에서 식별된 AI 시스템의 개인정보보호 위험을 조사하고 문서화한다.
	2.11 매핑 기능에서 식별된 공정성 및 편향을 평가하고 그 결과를 문서화한다.
	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 문서화한다.
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 구축한다.
	3.2 현재 가용 측정 기술을 사용하여 AI 위험을 평가하기 어렵거나 관련 지표를 아직 사용할 수 없는 경우 위험 추적 접근 방법이 고려된다.
	3.3 문제를 보고하고 시스템 결과에 이의를 제기하기 위한 최종 사용자 및 영향을 받는 커뮤니티의 피드백 프로세스를 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 정보를 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 결과를 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 문서화한다.

[신뢰할 수 있는 AI시스템 개발을 위한 필요기능] - 관리

범주(기능)	하위범주
관리 1 매핑 및 측정 기능으로부터 얻은 평가 및 기타 분석 결과를 기반으로 AI 위험에 대해 우선순위 부여, 대응하며, 관리한다.	**1.1** AI 시스템이 의도한 목적 및 목표를 달성했는지 여부와 시스템의 개발 또는 배포를 진행해야 하는지 여부에 대한 결정을 내린다.
	1.2 문서화된 AI 위험은 영향, 가능성, 가용 리소스 또는 방법에 따라 그 우선순위가 지정된다.
	1.3 매핑 기능을 통해 식별된 우선 순위가 높은 AI 위험에 대응하기 위한 방법을 개발, 계획 및 문서화한다. 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 문서화한다.
관리 2 관련 AI 행위자의 개입을 통해 AI 이점을 극대화하고 부정적인 영향을 최소화하기 위한 전략을 계획, 준비, 구현, 문서화하고 해당 정보를 제공한다.	**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 고려한다.
	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 절차를 준수한다.
	2.4 의도한 목적과는 다른 성능 또는 결과를 나타내는 AI 시스템을 대체, 해제 또는 비활성화하기 위한 메커니즘을 마련하고 관련 책무를 할당하고 파악한다.
관리 3 제3자 기관의 AI 위험 및 이점을 관리한다.	**3.1** 제3자 리소스의 AI 위험 및 이점을 정기적으로 모니터링하고 위험 제어를 적용하고 문서화한다.
	3.2 AI 시스템의 정기적 모니터링 및 유지 관리의 일환으로 개발용으로 사용되는 사전 학습된 모델을 모니터링한다.

범주(기능)	하위범주
관리 4 식별 및 측정된 AI 위험에 대해 위험 처리(대응 및 복구 포함) 및 커뮤니케이션 계획을 문서화하고 이를 정기적으로 모니터링한다.	**4.1** 배포 후 AI 시스템에 대한 모니터링 계획을 구현한다. 여기에는 사용자 및 기타 관련 AI 행위자의 의견을 수집하고 평가하기 위한 메커니즘, 이의 제기, 중단, 해제, 사고 대응, 복구 및 변경 관리가 포함된다.
	4.2 지속적인 개선 활동이 AI 시스템 업데이트에 통합되며, 여기에는 이해당사자(관련 AI 행위자 포함)와의 정기적인 참여가 포함된다.
	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 문서화한다.

[별첨] 유네스코 인공지능 윤리권고 (2021년 11월 25일)

유네스코 인공지능 윤리권고 (2021년 11월 25일)

(a) 인공지능 시스템은 현실 가상 환경에서 예측 및 의사 결정과 같은 결과를 도출하는 인지 과제의 학습 수행 능력을 생성하는 모델 및 알고리즘을 통합한 정보 처리 기술이다
(b) 인공지능 시스템 수명 주기를 관리, 운영, 매매, 재무, 모니터링 및 심사, 검증, 종료, 해체, 폐기를 비롯하여 연구, 설계, 개발로부터 배포, 사용에 이르는 과정까지 포함하는 것으로
이해할 때, 인공지능 시스템에 관한 윤리적 질문은 각 단계마다 존재한다
(c) 인공지능 시스템은 인공지능이 의사 결정, 고용 및 노동, 사회적 상호작용, 건강관리, 교육, 미디어, 정보 접근성, 디지털 격차, **개인정보** 및 소비자 보호, 환경, 민주주의, 법치주의, **보안** 및 치안유지, 군민 양용(dual use), 그리고 표현의 자유, 프라이버시, 비차별을 포함하는 인권 및 근본적 자유에 미치는 영향을 비롯하여 (단, 이에 국한되지 않는) 새로운 유형의 여러 윤리 문제를 유발한다.

. 본 권고는 인공지능 시스템 수명 주기 전반에 걸쳐 법적 규제적 틀을 개발하고 기업의 의무를 촉진시켜야 할 행위 주체이자 관계 당국인 회원국을 대상으로 한다. 또한, 본 권고는 **인공지능 시스템 수명 주기 내내 인공지능 시스템에 대한 윤리영향평가의 토대를 제공**함으로써 공공 민간 부문을 비롯한 모든 인공지능 행위 주체에게 윤리 지침 을 제공한다.

[별첨] 유네스코 인공지능 윤리권고 (2021년 11월 25일)

유네스코 인공지능 윤리권고 (2021년 11월 25일)

(a) 인공지능 시스템은 현실 가상 환경에서 예측 및 의사 결정과 같은 결과를 도출하는 인지 과제의 학습 수행 능력을 생성하는 모델 및 알고리즘을 통합한 정보 처리 기술이다
(b) 인공지능 시스템 수명 주기를 관리, 운영, 매매, 재무, 모니터링 및 심사, 검증, 종료, 해체, 폐기를 비롯하여 연구, 설계, 개발로부터 배포, 사용에 이르는 과정까지 포함하는 것으로
이해할 때, 인공지능 시스템에 관한 윤리적 질문은 각 단계마다 존재한다
(c) 인공지능 시스템은 인공지능이 의사 결정, 고용 및 노동, 사회적 상호작용, 건강관리, 교육, 미디어, 정보 접근성, 디지털 격차, **개인정보** 및 소비자 보호, 환경, 민주주의, 법치주의, **보안** 및 치안유지, 군민 양용(dual use), 그리고 표현의 자유, 프라이버시, 비차별을 포함하는 인권 및 근본적 자유에 미치는 영향을 비롯하여 (단, 이에 국한되지 않는) 새로운 유형의 여러 윤리 문제를 유발한다.

. 본 권고는 인공지능 시스템 수명 주기 전반에 걸쳐 법적 규제적 틀을 개발하고 기업의 의무를 촉진시켜야 할 행위 주체이자 관계 당국인 회원국을 대상으로 한다. 또한, 본 권고는 **인공지능 시스템 수명 주기 내내 인공지능 시스템에 대한 윤리영향평가의 토대를 제공**함으로써 공공 민간 부문을 비롯한 모든 인공지능 행위 주체에게 윤리 지침 을 제공한다.

Right to Privacy, and Data Protection
프라이버시 권리 및 정보 보호

32. Privacy, a right essential to the protection of human dignity, human autonomy and human agency, must be respected, protected and promoted throughout the life cycle of AI systems.

32. 인간 존엄성, 인간 자율성, 인간 활동의 보호에 핵심적 권리인 **프라이버시**는 인공지능 시스템 수명 주기 전 영역에서 반드시 존중, 보호, 증진 되어야 한다.

It is important that data for AI systems be collected, used, shared, archived and deleted in ways that are consistent with international law and in line with the values and principles set forth in this Recommendation, while respecting relevant national, regional and international legal frameworks.

인공지능 시스템을 위한 **데이터**가 국제법에 부합하고 본 권고에서 명시된 가치와 원칙에 맞게, 그리고 또한 유관 국가 지역 국제적 법률 틀을 존중하는 식으로 **수집, 사용, 공유, 보관, 삭제**되는 것은 중요하다.

33. Adequate data protection frameworks and governance mechanisms should be established in a multi-stakeholder approach at the national or international level, protected by judicial systems, and ensured throughout the life cycle of AI systems.

33. 적절한 **데이터 보호 틀 및 거버넌스 메커니즘**은 국가 국제적 차원의 다자적 접근법 으로 확립되어야 하고, 사법 체계에 의해 보호받아야 하며, **인공지능 시스템 수명 주기** 전 영역에서 보장되어야 한다.

Data protection frameworks and any related mechanisms should take reference from international data protection principles and standards concerning the collection, use and disclosure of personal data and exercise of their rights by data subjects while ensuring a legitimate aim and a valid legal basis for the processing of personal data, including informed consent.

데이터 보호 틀 및 관련 메커니즘은 **개인 정보의 수집 사용 공개 및 데이터 주체의 권리 행사와 관련된 국제 데이터 보호 원칙 기준을 참고**해야 하며, 동시에 인지동의를 비롯하여 개인 데이터 처리를 위한 적법한 목적 및 타당한 법적 토대도 확고히 해야 한다.

34. Algorithmic systems require adequate privacy impact assessments, which also include societal and ethical considerations of their use and an innovative use of the privacy by design approach.

34. 알고리즘 시스템은 이의 개인정보 사용과 설계 단계에서의 개인정보의 혁신적 사용에 대한 사회 윤리적 고려를 포함하여 적절한 프라이버시 영향평가를 요구한다.

AI actors need to ensure that they are accountable for the design and implementation of AI systems in such a way as to ensure that personal information is protected throughout the life cycle of the AI system.

인공지능 행위 주체는 가령 개인 정보가 인공지능 시스템의 전 영역에서 보호받음을 보장하는 식으로 그들이 인공지능 시스템의 설계 및 구현에 책임을 짐을 보장해야 한다.

Transparency and explainability
투명성 및 설명가능성

37. The transparency and explainability of AI systems are often essential preconditions to ensure the respect, protection and promotion of human rights, fundamental freedoms and ethical principles.
37. 인공지능 시스템의 **투명성 및 설명가능성**은 인권 및 근본적 자유, 윤리적 원칙에 대한 존중 보호 증진을 확고히 함에 있어 필수 선결조건이다.

Transparency is necessary for relevant national and international liability regimes to work effectively.
투명성은 관련 국내 국제 법적 책임 체제가 효과적으로 작동하는 데에 필수적이다.

A lack of transparency could also undermine the possibility of effectively challenging decisions based on outcomes produced by AI systems and may thereby infringe the right to a fair trial and effective remedy, and limits the areas in which these systems can be legally used.
또한 투명성의 부족은 인공지능 시스템이 산출한 결과물에 기반한 의사결정에 대한 **실질적 이의제기 가능성을 저해**할 수 있고, 이에 따라 공정한 재판 및 배상을 받을 권리를 침해 할 수 있으며, 이러한 시스템이 법적으로 사용될 수 있는 영역을 제한할 수 있다.

38. While efforts need to be made to increase transparency and explainability of AI systems, including those with extra-territorial impact, throughout their life cycle to support democratic governance, the level of transparency and explainability should always be appropriate to the context and impact, as there may be a need to balance between transparency and explainability and other principles such as privacy, safety and security.
38. 인공지능 시스템 수명 주기 전 영역에서 민주적 거버넌스를 지원하기 위해서는 (법역 외 영향까지도 포함하여) 이의 투명성 및 설명가능성을 향상시키려는 모든 노력이 이루어 져야 하는 반면, 투명성 및 설명가능성의 수준은 항상 해당 맥락과 영향 정도에 따라 적절한 수준이어야 하는데, 이는 투명성 및 설명가능성과 프라이버시, 안전 및 보안과 같은다른 원칙 사이에서 균형을 이루어야 할 필요성이 있을 수 있기 때문이다

Transparency and explainability
투명성 및 설명가능성

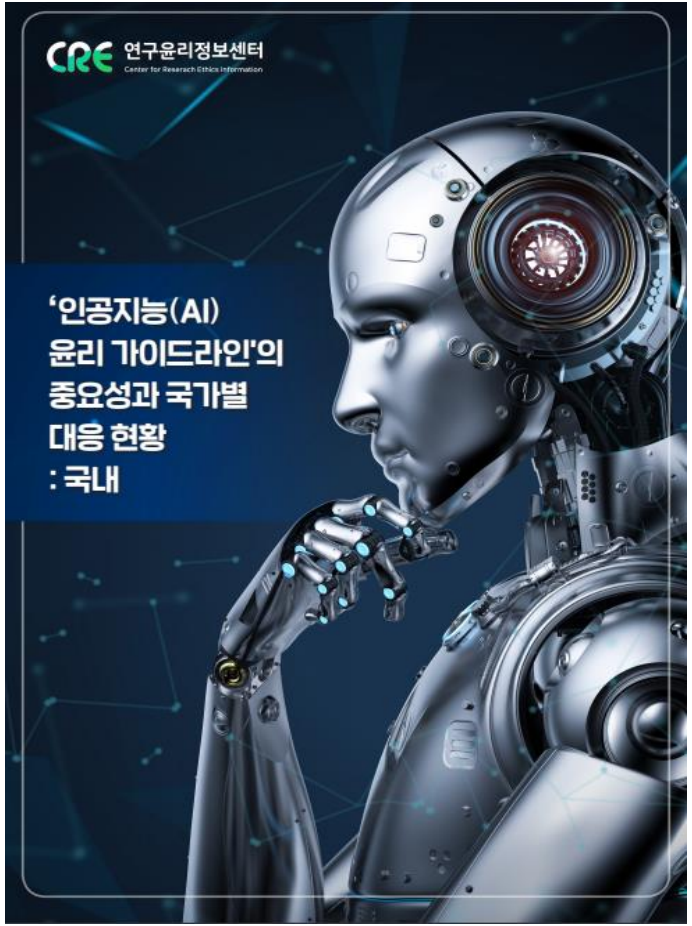
People should be fully informed when a decision is informed by or is made on the basis of AI algorithms, including when it affects their safety or human rights, and in those circumstances should have the opportunity to request explanatoryinformation from the relevant AI actor or public sector institutions.
사람들은 의사결정이 인공지능 알고리즘으로부터 정보를 얻거나 이에 기반하여 내려지는 경우, 특히 이로써 그들의 안전 또는 인권에 영향이 가는 경우에 사전에 알 수 있어야 하고, 그러한 상황에서 관련 인공지능 행위 주체 또는 공공 기관에서 **정보 설명을 요구할 수 있는 기회**를 가지고 있어야 한다.
In addition, individuals should be able to access the reasons for a decision affecting their rights and freedoms, and have the option of making submissions to a designated staff member of the private sector company or public sector institution able to review and correct the decision.
더욱이, 개인은 자신의 권리 및 자유에 영향을 미치는 **결정에 대한 근거에 접근할 수 있어야 하고, 이 결정을 검토 정정**할 수 있는 민간 기업 또는 공공 기관의 담당 직원에게 **의견을 개진할 수 있는 선택권**을 가지고 있어야 한다.
AI actors should inform users when a product or service is provided directly or with the assistance of AI systems in a proper and timely manner.
인공지능 행위 주체는 제품이나 서비스가 직접적으로 또는 인공지능 시스템의 보조를 통해 제공될 때 사용자에게 이를 시기적절하게 **사전 통보**해야 한다.

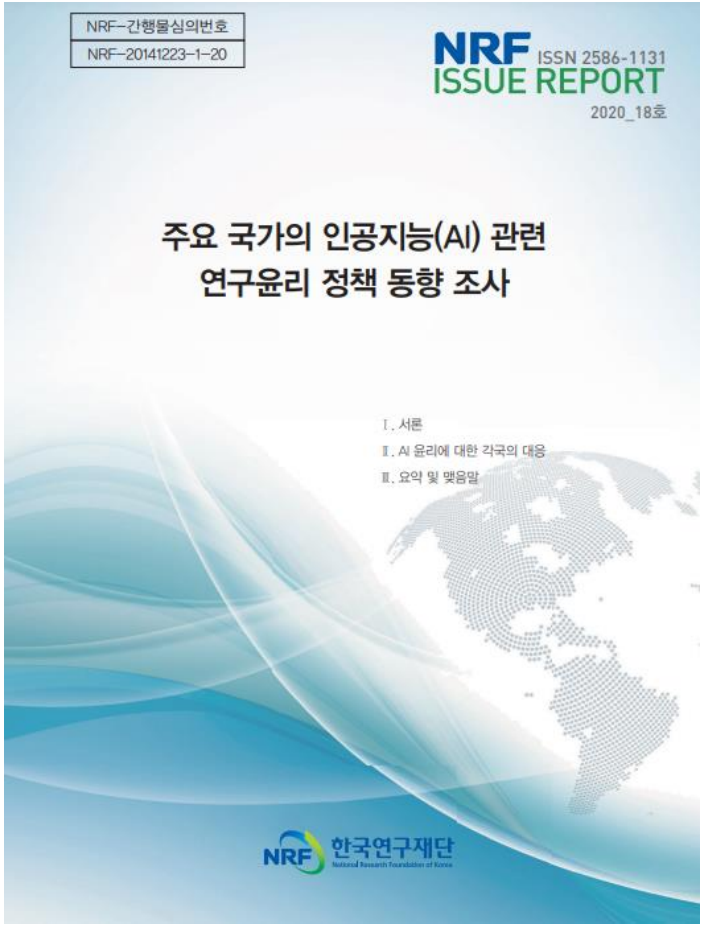
Responsibility and accountability
책임 및 책무

42. AI actors and Member States should respect, protect and promote humanrights and fundamental freedoms, and should also promote the protection of the environment and ecosystems, assuming their respective ethical and legal responsibility, in accordance with national and international law, in particular Member States' human rights obligations, and ethical guidance throughout the lifecycle of AI systems, including with respect to AI actors within their effective territory and control.
42. 인공지능 행위 주체 및 회원국은 인공지능 시스템 수명 주기 내내 국내 국제법, 특히 국제 회원국의 인권 준수 의무와 인공지능 시스템의 수명 주기 전 영역에서의 윤리 지침에 따라 인공지능 행위 주체를 효과적인 영역과 통제 하에 둠으로써, 각각의 윤리적 법적 책임를 지고 인권 및 근본적 자유를 존중 보호 증진하며 환경 및 생태계 보호를 장려해야 한다.
The ethical responsibility and liability for the decisions and actions based in any way on an AI system should always ultimately be attributable to AI actors corresponding to their role in the life cycle of the AI system.
일단 인공지능 시스템에 기반하여 내린 결정 및 행동에 대한 윤리적 책무와 법적 책임은 항상 궁극적으로 인공지능 시스템 수명 주기에서의 해당 역할의 인공지능 행위 주체에게 귀속된다.
43. Appropriate oversight, impact assessment, audit and due diligence mechanisms, including whistle-blowers' protection, should be developed to ensure accountability for AI systems and their impact throughout their life cycle.
43. 인공지능 시스템과 이의 수명 주기 전 영역에서의 영향력에 대한 책임을 보장하기 위해서는 제보자 보호를 비롯하여 **적절한 감독, 영향 평가, 감사, 실사 메커니즘이 개발**되어야 한다.
Both technical and institutional designs should ensure auditability and traceability of (the working of) AI systems in particular to address any conflicts with humanrights norms and standards and threats to environmental and ecosystem well-being.
특히 인권 개념 기준과의 충돌 및 환경 및 생태계의 안녕에 대한 위협이 발생하는 경우, 기술적, 제도적 설계는 이를 해결할 수 있도록 인공지능 시스템(또는 이의 작동)에 대한 **감사가능성 및 추적가능성**을 보장해야 한다.

[별첨] 한국연구재단- 연구윤리정보포탈 자료

한국연구재단 CRE 연구윤리정보포털 https://www.cre.or.kr/





AI 윤리 대응 동향

■ 네이버

○ 인공지능 기반 서비스에서 안전한 데이터 활용을 위해 'Privacy By Design 원칙'을 적용

- 서비스 출시 단계에서부터 높은 프라이버시 보호 기준으로 데이터 활용에 있어 법령 위반사항이나 이용자 프라이버시 보호 관점 에서 문제가 없는지를 검토하고 있다.

○ '네이버 프라이버시센터' 홈페이지, 개인정보 보호 블로그 및 SNS를 통해 이용자 프라이버시 보호를 위한 활동과 관련 정보를 공개하고 이용자의 개선제안 등을 받고 있다.

'네이버 프라이버시 센터' 홈페이지에서는 네이버의 개인정보 보호원칙), 서비스 운영 관련 정책, 보호 활동, 투명성 보고서, 개인정보 보호와 활용 논의를 위한 Privacy White Paper 등 프라이버시 관련 보고서를 제공하고 있다

| 네이버의 프라이버시 관련 보고서 |

구분	내용	발행주기
투명성 보고서 통계	네이버(주)에서 수사기관에 제공한 이용자 정보 통계	연 2회
개인정보보호 리포트	네이버(주)의 개인정보 보호 활동	연 1회
Privacy White Paper	'이용자 프라이버시 보호' 관련 전문 연구 수행 결과	연 1회

자료: 네이버 프라이버시센터 홈페이지

■ 카카오

○ 2018년 1월 31일, 국내 업계 최초로 인공지능 알고리즘 윤리 헌장을 발표했다.

| 카카오의 윤리 헌장 |

원칙	내용
1. 카카오 알고리즘의 기본원칙	카카오는 알고리즘과 관련된 모든 노력을 우리 사회 윤리 안에서 다하며, 이를 통해 인류의 편익과 행복을 추구한다.
2. 차별에 대한 경계	알고리즘 결과에서 의도적인 사회적 차별이 일어나지 않도록 경계한다.
3. 학습 데이터 운영	알고리즘에 입력되는 학습 데이터를 사회 윤리에 근거하여 수집·분석·활용한다.
4. 알고리즘의 독립성	알고리즘이 누군가에 의해 자의적으로 훼손되거나 영향받는 일이 없도록 엄정하게 관리한다.
5. 알고리즘에 대한 설명	이용자와의 신뢰 관계를 위해 기업 경쟁력을 훼손하지 않는 범위 내에서 알고리즘에 대해 성실하게 설명한다.

■ 삼성전자

○ 2018년 국내기업으로는 최초로 인공지능 국제 컨소시엄인 PAI(Partners on AI)에 가입

- 인공지능 미래에 대한 범사회적인 논의를 진행 중인 삼성전자는 향후

AI 안전성(Safety-Critical AI)과 AI 공정성·투명성·책임성(Fair, Transparent, and Accountable AI), AI의 사회적 영향(Social and Societal Influences of AI) 등 다양한 분야에 참여를 확대할 계획이다.

출처 : https://ethics.moe.edu.tw/files/resource/ebook/file/ebook_01_kr.pdf

■ 한국인공지능협회

○ 2012년 스타트업 실무자들의 커뮤니티로 시작하여 2017년 사단법인으로 설립. "산업의 지능화"와 "AI 기술의 대중화"를 목표로 인공지능 기술 기업 및 유관기관들을 중심으로 네트워크를 구축

○ 인공 지능 산업에 필요한 윤리의식과 교육의 필요성을 논의하기 위한 목적으로 2019년 7 월 '제1회 인공지능 윤리+교육 포럼'을 개최하였으며, 포럼 이후 메니패스토의 일환으로 협회 클러스터 기업 72개 업체가 참여한 2019 인공지능 윤리+교육 포럼 공동선언문을 채택했다.

| 2019 인공지능 윤리+교육 포럼 공동선언문 |

본 선언문은 대한민국 인공지능 기술의 지속적 성장과 안전한 발전을 위한 윤리적 연구 개발 뿐 아니라, 이와 연계하여 올바른 인공지능 교육을 할 수 있도록 개인과 기업, 국가 그리고 세계기구에 실천적 가이드라인을 제시하는 데 그 목적이 있다.

1. 인공지능 기술은 처음부터 '사람 우선' 원칙으로 인류에게 육체적·정신적·환경적으로 도움을 줄 수 있도록 연구·개발되어야 하고 동시에 설명될 수 있어야 한다. 또한, 우리가 그것을 신뢰하고 공정하게 공유하여 인류 공동의 선을 실현할 수 있도록 해야 한다.

2. 국민 개인은 '자기의 기술(The Technology of the Self)'을 학습하여 개인 데이터를 스스로 관리해야 한다. 또 미래 인공지능 기술이 가져올 직업적 변화에 대비하여 자신의 적성과 능력을 파악, 소비자인 동시에 생산자가 될 수 있는 새로운 디지털기술을 습득하고 그것을 발전시켜야 한다.

3. 모든 기업은 시작단계에서부터 '안전 우선'을 원칙으로 관계자 교육훈련을 포함한 인공지능 개발과정을 설계하고, 해당 제품의 사람과 사회에 대한 영향을 종합적으로 분석하여 차별과 편견이 없도록 제작해야 한다. 그리고 비상시를 대비하여 위험요소를 즉각 제거할 수 있는 통제기술을 해당 제품에 포함해야 하며, 제품의 전체 제작과정을 표준화시키고, 사후 관리에 책임을 지며, 영업비밀을 훼손하지 않는 범위에서 그것을 설명할 수 있어야 한다.

4. 국가 교육기관은 국민 각자가 국민공동기본교육과정을 통해 생애주기별로 필요한 '스마트 라이프(Smart Life)' 활용기술을 습득하여 일상 생활능력을 확장할 수 있게 하며, 직업적으로 개인의 역량을 최대로 향상시킬 수 있는 '자기 계량화(Quantified Self)' 기술을 포함한 '인간공학(Human Engineering)'에 대한 학습기회를 제공해야 한다.

5. 국가는 국민의 인권과 재산, 프라이버시를 보호하기 위해 산업별 인공지능 기술의 긍정적·부정적 영향을 전체적으로 조사·분석·예측하여 그 결과를 공유하고, 관계자에 대한 자격부여 및 교육·훈련을 해야 한다. 특히, 종합적 제품관리를 위해 기업 경쟁력을 침해하지 않는 범위에서 산업별 효준화된 평가 매트릭스를 적용하여 투명성과 타당성을 확보하고, 합리적 개입과 감독을 포함한 국가 차원의 가버넌스시스템을 구축해야 한다.

6. 인공지능 기술을 그 경계가 없이 세계적으로 확장될 수 있으므로 개인과 기업은 세계시민으로서의 역할교육을 받아야 하며, 한 나라의 국가적 가브넌스 시스템은 UN과 같은 세계기구의 글로벌 시스템과 연동시켜 지속해서 발전해 나가야 한다.

7. UN등 세계기구는 인공지능 기술이 인간의 생리적 진화를 앞지를 수 있는 문명사적 변혁에 대비하여, 인간의 존엄성과 그 존재론적 의미와 가치를 재발견해야 한다. 나아가서는 '사람 우선' 의 원칙을 기반을로 인공지능과 인간의 표준화된 기술적 협업 방법을 개발하고 보급하여 상호갈등을 없애고, 세계평화와 질서를 유지할 수 있는 정책을 강구해야 한다.

자료: 한국인공지능협회 홈페이지

■ 고등과학원 초학제연구단

■ 한국인공지능법학회

■ 한국인터넷윤리학회

■ 정보통신정책연구원

■ 방송통신위원회

○ 2019년 11월 11일 정보통신정책연구원과 함께 AI 시대 정부, 기업, 이용자 등 구성원이 지켜야 할 '이용자 중심의 지능정보서비스 기본 원칙' 발표.

○ 지능정보서비스 기본 원칙 주요내용
- (**사람 중심**의 서비스 제공) 지능정보서비스의 제공과 이용은 사람을 중심으로 그 기본적 자유와 권리를 보장하고 인간의 존엄성을 보호할 수 있는 방향으로 이루어져야 한다.
- (**투명성과 설명가능성**) 지능정보서비스가 이용자에게 중대한 영향을 끼칠 경우 기업의 정당한 이익을 침해하지 않는 범위에서 이용자가 이해할 수 있도록 관련 정보를 작성해야 하며, 이용자 기본권에 피해를 유발했을 때 예측, 추천, **결정의 기초로 사용한 주요요인을 설명**할 수 있어야 한다.
- (**책임성**) 지능정보사회의 구성원들은 지능정보서비스의 올바른 기능과 사람 중심 가치의 보장을 위한 공동의 책임을 인식하고, 관련한 법령과 계약을 준수한다.
- (**안전성**) 안전하고 신뢰 가능한 지능정보서비스의 개발과 이용을 위해 모두가 노력하고, 지능정보서비스가 초래할 수 있는 **피해에 대한 자율적인 대비체계**를 서비스 제공자와 이용자가 수립하고 운영한다.
- (**차별금지**) 지능정보서비스가 사회적·경제적 불공평이나 격차를 초래할 수 있다는 점을 인식하고, 알고리즘 개발과 사용의 모든 단계에서 차별적 요소를 최소화할 수 있도록 노력한다.
- (**참여**) 지능정보사회의 구성원들은 공적인 이용자 정책 과정에 차별 없이 참여할 수 있으며, 공적 주체는 제공자와 이용자가 실질적으로 의견을 제시할 수 있는 정기적인 통로를 조성해야 한다.
- (**프라이버시와 데이터거버넌스**) 지능정보서비스의 개발, 공급 및 이용의 전 과정에서 **개인정보 및 프라이버시를 보호**하며, 구성원들은 기술적 이익의 향유와 프라이버시 보호 사이의 균형을 위해 지속적인 의견 교환에 참여한다.

○ 이용자 보호를 위한 공동의 노력
- 지능정보사회의 구성원들은 공동의 기본 원칙에 입각하여 지능정보사회의 기본 가치를 수호할 수 있도록 자율적인 노력을 지속한다.
- 지능정보사회 이용자 보호를 위한 실질적이고 지속적인 논의를 위해 지능정보사회의 구성원들이 참여하는 **협의회를 구성하여 운영**한다.

출처 : https://ethics.moe.edu.tw/files/resource/ebook/file/ebook_01_kr.pdf

■ 정부

○ 2019년 12월 17일 정부는 문재인 대통령 주재로 열린 제 53회 국무회의에서, 과학기술정보통신부를 비롯한 전 부처가 참여하여 마련한「인공지능(AI) 국가전략」을 발표

▶ AI 반도체 세계 1위, 전국 단위 AI 거점화 등 세계를 선도하는 AI 생태계 조성

▶ 전 생애·모든 직군에 걸친 AI교육 실시 및 세계 최고의 AI인재 양성

▶ 현 전자정부를 차세대 지능형 정부로 대전환하여 국민 체감도 향상

▶ 사회보험 확대 등 일자리 안전망 확충 및 AI 윤리 정립으로 사람 중심 AI 실현

- AI를 통해 경제효과 최대455조원 창출, 삶의 질 세계 10위 도약(~'30) -

| AI 윤리 규범 및 가이드라인 마련 현황 |

○ 역기능 방지 및 AI 윤리체계 마련

❶ AI 기반 사이버침해 대응체계 고도화('20~)

❷ 딥페이크* 등 신유형의 역기능 대응을 위한 범부처 협업체계 구축('20)

(AI 기반 영상 합성기술 또는 그 영상. 신시장 창출과 동시에 명예훼손 등 부작용도 우려)*

❸ AI 신뢰성·안전성 등을 검증하는 품질관리체계 구축 추진('20~)

❹ OECD 등 글로벌 규범에 부합하는 AI 윤리기준 확립('20) 및 AI 윤리교육 커리큘럼* 개발·보급('21~)

((학생·이용자) AI와 생명윤리, 개인정보보호 / (개발자) 윤리적 AI 설계, 정보보안 등)*

❺ 이용자 보호를 위한 중장기적 정책 수립 지원체계 마련

○ 정부는 대통령 직속의 현 4차산업혁명위원회를 AI의 범국가 위원회로 역할을 재정립하여 이번 전략의 충실한 이행을 위한 범정부 협업체계를 구축해 나갈 방침

- 특히, 대통령 주재 전략회의를 개최하여 전 국민 교육, 전 산업 AI 활용 등 범정부적 과제의 실행력을 확보하고, 대국민 성과 보고대회도 병행하여 국민의 참여와 성과 확산에도 노력할 계획

■ 금융위원회

○ 금융 분야 인공지능(AI) 윤리 가이드라인 및 인프라를 조성해 AI 기술 활성화에 나설 계획.

- 금융위원회는 2020년 7월 16일 국내·외 금융 분야 AI 활용 및 정책 동향 파악을 목표로 '금융 분야 인공지능(AI) 활성화' 워킹 그룹을 조직, 첫 회의를 개최

금융 분야 AI 활성화 정책을 추진할 수 있도록 워킹 그룹을 구성하고, 약 4개월간 운영해 '금융 분야 AI 활성화 방안'을 마련할 계획이다.

워킹 그룹은 ▲AI 관련 규제 ▲AI 인프라 ▲소비자 보호 ▲레그테크*와 섭테크* 접목 4개 측면에서 AI 활성화 정책을 추진한다.

○ AI 활성화를 목표로 관련 규제 개선 및 규율 체계를 정립.

- 가이드라인 형태로 AI 금융 서비스 개발에 특화한 실무 프로세스를 마련하고, 과학기술정보 통신부 AI 법제정비단과 협력해 적법성·공정성을 담은 '금융 분야 AI 윤리 가이드라인'을 갖춘다.

○ AI 활용 시 소비자 권한을 보호할 수 있는 체계를 확보.

- AI 업무 처리로 발생하는 소비자 피해를 해결할 수 있도록 책임 주체와 구제 절차 등 기준을 마련한다. AI가 도출한 결과를 객관적으로 설명할 수 있는 '설명가능 AI(XAI)' 기준도 정립한다.

*(*레그테크는 '레귤레이션(규제)'과 기술을 의미하는 '테크놀리지(기술)'의 합성어로 AI를 활용해 복잡한 금융 규제를 기업이 쉽게 이해하고 지킬 수 있도록 하는 기술이다. *섭테크는 '슈퍼비전(감독)'과 '테크놀리지(기술)'의 합성어로 신기술을 활용해 금융 감독 업무를 효율적으로 수행하는 기술이다.)*

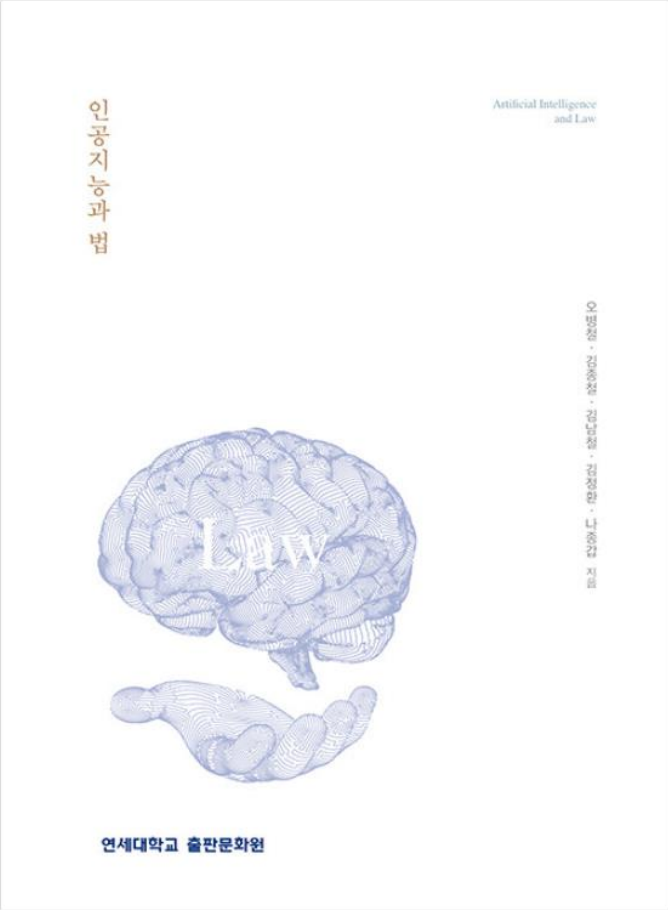
■ 과학기술정보통신부

○ AI 법제도 개선 및 윤리 정립

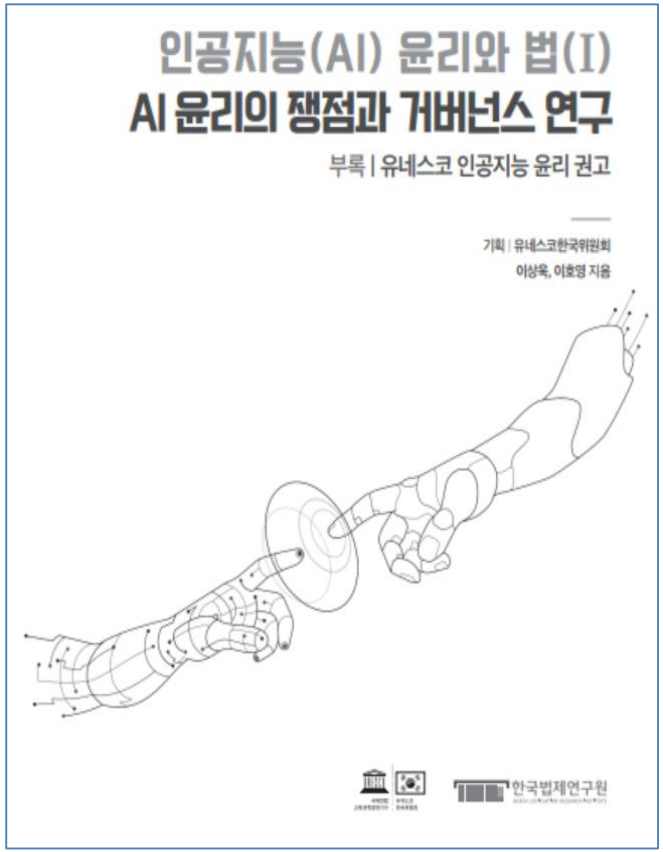
향후 계획으로는 AI로 인한 경제·사회 전반의 패러다임 변화에 대응하고자 규제개선 사항을 종합하여 AI분야 법제도 개선 로드맵을 제시할 계획이다.('20.11월 잠정) 아울러 금년 12월 AI시대의 기본법제인 『지능정보화 기본법』이 시행됨에 따라 하위법령을 완비하고, 주요국, 국제기구 등의 AI 윤리규범을 비교 분석하여 우리나라의 AI 윤리 기준도 정립할 예정이다.

출처 : https://ethics.moe.edu.tw/files/resource/ebook/file/ebook_01_kr.pdf

[별첨] 인공지능과 법



[별첨] 유네스코 인공지능 윤리권고 (2021년 11월 25일 193개 유네스코 회원국의 만장일치로 'AI 윤리 권고'를 공식적으로 채택)



FPRI
Future Policy Research Institute

news1 KOREA

"인공지능 판사, 민주적 정당성 차원서 문제될 수 있어"

서미선 2019.12.18. 18:17

인공지능(AI) 판사가 도입돼 국가 권력을 AI가 행사할 경우 민주적 정당성 차원에서 문제가 될 수 있다는 의견이 나왔다.
김중권 중앙대 법학전문대학원 교수는 18일 서울 서초동 서울법원종합청사 청심홀에서 열린 'AI와 법 그리고 인간' 심포지엄에서 주제 발표를 통해 이같이 밝혔다.
김 교수는 "민주주의와 법치국가원리는 민주적 정당성을 지닌 자연인에 의한 지배를 바탕으로 한다"며 "국가권력 행사에서의 AI도입은 특히 인적 민주적 정당성 차원에서 문제될 수 있다"고 지적했다.
그는 "기계가 AI에 의해 스스로 향상될 수 있는 시점까지 어떤 전문가시스템도 인간통제 없이 출현할 수는 없다"며 "인풋과 아웃풋을 통제할 수 있는 전문가를 항상 필요로 할 것이고, 정보지식을 구비한 더 유능한 법조인이 요구된다"고 봤다.
이어 "맹목적 기술신봉에서 법조직업의 종말을 외치는 건 전혀 타당하지 않다"며 "현재의 한계를 인식하고 AI프로그램을 효과적으로 활용해야 할 것"이라고 언급했다.
그러면서 "AI 입법과 AI 재판, AI 행정에 관한 움직임 이면엔 기성 국가작용 메커니즘에 대한 심각한 불신이 있다"고 진단했다.

2019.12.18
AI와 법 그리고 인간

https://www.msn.com/ko-kr/news/national인공지능-판사-민주적-정당성-차원서-문제될-수-있어/ar-BBY6UoD

사법분야에서 AI를 활용하려면 윤리적 사용지침이 마련돼야 한다는 제언도 나왔다.
에스토니아의 Kai Härmand 법무부차관(판사)은 450여개 기관과 150여개 공공기관이 쓰는 에스토니아의 국가정보 교환플랫폼인 엑스로드(X-road)를 소개하고, **사법분야에선 검색·번역·기록·자문영역에서 AI를 활용할 수 있다면서 이처럼 말했다**.
오스트리아의 마르크 코켈버그(Mark Coeckelbergh) 빈 대학교 교수는 "**AI로 인한 도덕적·법적 책임주체는 기술이 될 수 없고 인간이 돼야 하는데, 주체가 불분명하고 AI 알고리즘이 불투명한 문제가 있다**"고 지적하며 AI 윤리지침을 제시했다.
구체적으로 **인간의 감독과 기술적 안정성, 데이터 관리, 투명성, 차별금지, 공정성이다.**
옥스퍼드 딥 테크 분쟁해결연구소 리 지 엔(Ji En Lee) 연구원은 증거분석과 자동기록을 통해 판사업무를 지원하는 중국의 '206' 시스템, 교통사고 사건을 위한 싱가포르의 시뮬레이션 프로그램, 범죄자 위험평가 도구인 미국의 콤파스(COMPAS)를 소개했다.
이어 "법률분야에 AI를 적용할 때 생길 수 있는 윤리적 문제를 해결하려면 공정, 투명, 설명 가능성, 기본권 존중, 데이터 정확성과 보안, 협력과 포용, 이용자에 의한 통제 등 요소를 고려해 AI를 개발해야 할 것"이라고 말했다.
발표에 앞서 강현중 사법정책연구원장은 개회사에서 "이미 전산화돼 축적된 법률문서 및 정보가 AI에 의해 활용된다면 법조인 업무에도 큰 변화가 올 것"이라며 "정보기술(IT) 강국인 한국은 AI영역에서도 앞서나갈 저력이 있다"고 밝혔다.
조재연 법원행정처장은 축사에서 법원이 추진하는 차세대전자소송 시스템구축 사업을 언급하고 "축적된 기존 전자소송문서 등 정보를 빅데이터 형태로 AI기술에 활용할 수 있을 것"이라고 설명했다.
이번 심포지엄은 사법정책연구원이 중앙대 인문콘텐츠연구소, 한국인공지능법학회, AI정책포럼과 공동으로 대한변호사협회와 한국연구재단, 교육부 후원을 받아 개최했다.

[별첨] AI 윤리

□ 윤리 : "사람으로서 마땅히 행하거나 지켜야 할 도리"

ethic : "A set of moral principles, especially ones relating to or affirming a specified group field, or form of conduct"

□ 윤리적 합의 : 개인/기업/사회/법적 가치의 지향점을 고려하여 핵심존중가치를 균형있게 존중하는 방식으로 AI를 개발/활용하기 위한 합의적 절충점을 제도적 장치로 마련 (Trade Off)

□ 인공지능 윤리 : 인공지능 관련 이해관계자들이 준수해야 할 보편적 사회 규범

AI의 ''자동화된 결정(automated decisions)'이 다양한 사회적 가치를 최대한 존중하는 방식으로 활용되기 위해, 고민해야 할 주제를 고르고, 어떻게 제도화 해야 하는지 대한 논의가 필요하며, 이는 AI 개발/활용 전 과정에 적용되어야 한다.
* 관련주제 : 공정, 책임성, 투명성, 설명가능성, 자율/통제정도, 인권 대 AI의 법인격(인간중심주의 균형점), 기술혁신과 사화가치보전사이의 균형(AI기술개발금지논의), 윤리영향평가 필요여부

'인간중심적(human Centered)' (AI는 도구, AI의 결정에 대한 책임은 인간에게 있음)

- 자율주행과 같은 경우 인간의 실시간 개입이 어려운 경우 의사결정 과정에서 인간이 반드시 필수불가결한 방식으로 참여하는 Human in the Loop 방식이 아니라 인간이 어떤 방식이든지 AI의 설계와 작동 등 전체 관리 체계에서 중요한 역할을 담당하고 문제가 있다고 판단될 때는 전체 과정을 중지시킬 권한을 갖는 Human on the Loop 방식이 필수적임

□ 인공지능 거버넌스 (AI리스크에 기반한 AI 윤리의 제도화)

: 적응적 거버넌스(adaptive governance)라는 개념으로 접근할 필요

- 사회적 공감대와 합의가 도출된 영역부터 제도화
- 이후 기술발전, 사회적 인식 변화가 지속 반영되어야 함

정책수립(윤리영향평가, 거버넌스, 감독의무, 시스템 보안, 프라이버시 데이터 보호 등) → 모니터링 및 평가

AI-Ethical Impact Assessment (AI-EIA)

□ AI윤리의 법제화

: '적응적' 거버넌스를 구체적으로 어떻게 실현할 지 구체적 법제화 //TODO

인공지능은 **"인간(조직)"**이 제공한 데이터와 알고리듬에 기반하여 판단하므로,
→ 기본적으로 AI윤리원칙(AI Ethical Principles)에 대한 사회적 합의 및 이를 준수하려는 자율적 활동이 필요하고,
→ 더 나아가 최소한의 내용은 법적규제사항으로 정리되어야 한다.

자율성을 가진 AI에 대한 논의 필요, 책임 귀속의 최종 주체는 인간/사회가 되므로 개발/적용/규제 이전에 사회적 합의가 필수적임, 그래서 인공지능 윤리선언이 선행된 후 이를 기반으로 법제도영역에 반영하는 순서로 진행되어야 함.

AI윤리준수 점검리스트

<해외 주요 인공지능 윤리 체크리스트 현황>

	발표주체	체크리스트	
기업	Microsoft	AI 공정성 체크리스트 AI Fairness Checklist	'20.03
	Lex Mundi & Cambrian Futures	법인 법률업무를 위한 AI 준비도 체크리스트 AI Readiness Checklist for Corporate Legal Functions	'20.04
공공	WEF & 싱가포르 IMDA	AI 거버넌스 프레임워크 지침 Companion to the Model AI Governance Framework	'20.01
	EU	신뢰할 수 있는 AI를 위한 자율평가목록 The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment	'20.07
	영국 ICO	AI 및 데이터보호 안내 지침 Guidance on Artificial Intelligence and Data Protection	'20.07
학계	Carnegie Mellon University's Software Engineering Institute	윤리적 AI 경험 설계 체크리스트 Designing Ethical AI Experiences: Checklist and Agreement	'19.12

자료: 정보통신정책연구원 지능정보사회정책센터 내부자료(2020)

“만약 당신이 미래를 꿈꾸지 않거나 지금 기술개선을 위해 노력하지 않는다면 그건 곧 낙오되고 있는 것이나 마찬가지 입니다.”

그윈 쇼트웰(Gwynne Shtwell, SpaceX CEO, COO)

감사합니다

(facebook.com/sangshik, mikado22001@yahoo.co.kr)

[TODO]

정부의 '사람이 중심이 되는 인공지능 윤리기준'(2020)
카카오 알고리즘 윤리 헌장(2018)
네이버AI윤리준칙(2021)

지능정보화 기본법
금융 분야 AI 윤리 가이드라인