THE AI REPORT

美 NIST「AI 위험관리 프레임워크(AI RMF) 1.0」 분석 및 시사점

2023

NIA AI·Future Strategy Center

▶ 발행인 : 황 종 성

▶ 작 성
- 한국지능정보사회진흥원 정책본부 AI·미래전략센터
  김태순 책임연구원(ts_kim@nia.or.kr)

美 NIST 「AI 위험관리 프레임워크(AI RMF)」 분석 및 시사점

"AI 위험관리 프레임워크는 AI혁신을 가속화하고 개인의 권리, 자유, 형평성을 제한하거나 훼손시키지 않고 발전해야한다" - Don Graves 상무부차관

"AI 위험관리 프레임워크는 모든 분야와 규모의 기업, 조직이 AI 위험관리 접근 방식을 시작하거나 개선하는데 도움이 될 것" - Laurie E. Locascio. 표준기술부 장관 및 NIST 소장

1. AI 위험관리 프레임워크(AI Risk Management Framework) 개요

○ 미국 국립표준기술연구소(이하 NIST)는 18개월간의 연구를 통해 AI 시스템을 설계 · 개발 · 도입하는 조직을 위한 자발적 활용 가이드 문서인 'AI 위험관리 프레임워크 1.0(이하 AI RMF)'을 발표('23.1.26)

- **(추진근거)** '20년에 제정된 '국가인공지능 이니셔티브법(National Artificial Intelligence Initiative Act of 2020 (P.L. 116-283))'*에 따른 미 의회의 지시로 AI RMF를 수립

  * 공공 및 민간 영역에서 신뢰할 수 있는 AI 시스템 개발 등을 위해 국가 AI 이니셔티브를 수립·이행하는 내용을 포함

- **(목적)** 모든 분야 · 규모의 기업 · 조직이 AI 위험을 해결할 수 있도록 유연하고 체계적이며 측정 가능한 프로세스를 제공하여 AI 기술의 이점을 극대화 하고 동시에 개인 · 그룹 · 조직 · 사회에 부정적인 영향을 미칠 가능성을 줄이기 위해 수립

- **(주체 및 절차)** NIST 주관으로 정부 기관 · 민간기업 · 학계 · 비영리단체 · 시민 · AI 실무자 등 민관협력으로 정보제공요청(RFI)*, 3번의 워크숍, 2개의 초안 문서에 대한 공개 의견 수렴, 공개 포럼 등을 통해 수립

  * 정보제공요청(RFI, Request for information) : 정책 · 전략 · 프로젝트 기획 단계에서 전문가와 대중을 대상으로 필요한 정보를 공개적으로 요청하는 과정

○ AI RMF는 조직의 크기와 형태에 상관없이 조직의 전체적인 위험관리 정책에 포함되어 AI 시스템의 전 주기에서 신뢰할 수 있는 AI 시스템 구현을 위한 핵심 기능의 위험관리 프로파일을 제시

- **(적용대상)** 모든 분야 · 규모의 기업 · 조직의 'AI 시스템'

- **(적용주체 및 방식)** 'AI 행위자'*가 제시된 항목에 따라 자발적으로 참고 및 점검

  * 설계자·개발자·배포자·사용자, 고위 경영진이나 관리자 등

- **(적용시점)** AI 시스템의 설계, 개발, 활용, 테스트, 평가 등 모든 단계

- **(핵심내용)** 안전성, 책임·투명성 등 신뢰할 수 있는 AI 시스템의 7가지 특성을 제시하고, 이를 위한 조직의 핵심 기능을 4가지 영역(거버넌스, 매핑(위험 식별), 측정, 관리)에서 제시

○ AI RMF는 위험 기반으로 조직 · 기업 등에서 자발적 · 보편적으로 활용할 수 있도록 쉽게 쓰고, 모든 이해관계자가 자유롭게 참여하여 기존의 정책 · 표준 · 지침 등에 맞춰 수립한 것이 주요 특징

① **(협업과 참여)** 공개적이고 투명한 프로세스를 통해 수립 및 갱신해야하며 모든 이해당사자가 AI RMF 수립에 참여할 수 있는 기회를 부여

② **(보편적 활용)** 고위 경영진, 정부 관리, 비정부 단체 경영진, 비AI 전문가를 포함한 광범위한 대상이 이해할 수 있는 명확하고 쉽게 쓰며 동시에 실무자에게 유용한 기술적 정보를 제공

* AI 위험에 대한 분류, 용어, 정의, 지표 및 특성을 제공

③ **(조직의 전체적 위험관리에 포함)** AI RMF는 단독으로 활용할 것이 아니라 조직에서 추진하고 있는 기존 위험관리 전략 및 프로세스의 일환으로 포함되어 직관적이고 쉽게 적용해야 함

④ **(결과 중심적)** 획일적인 요구사항을 규정하기보다는 결과와 접근 방식의 범주를 제공

⑤ **(기존 체제 내에서 적용)** AI 위험관리에 대한 기존 표준 · 지침 · 모범 사례 · 방법 및 도구를 활용하고 추가로 개선된 리소스의 필요성을 설명해야 하며 국내 · 국제 법률 또는 규제 하에서 적용될 수 있어야함

⑥ **(빠른 업데이트)** AI의 신뢰성 및 활용에 대한 기술 · 인식 · 접근방식이 변화하고 이해관계자들이 AI RMF를 통해 학습함에 따라 신속히 업데이트 필요

○ 본 보고서에서 '23년 1월 미국 상무성 산하 국립표준기술연구소(NIST)에서 발표한 'AI 위험관리 프레임워크 1.0(AI RMF)'을 분석하고, 시사점을 도출

AI RMF에서 정의하는 용어

용 어	정 의	참고
AI 시스템	• 주어진 목표 설정에 따라 실제 또는 가상 환경에 영향을 미치는 예측, 권장 사항, 또는 의사 결정과 같은 결과물을 생성할 수 있는 엔지니어링 또는 기계 기반 시스템 ※ 다양한 수준의 자율성을 가지고 작동하도록 설계한 시스템	OECD Recommendation on AI(2019) ; ISO/IEC 22989(2022)
AI 행위자 (AI actors)	• AI를 배포하거나 운영하는 조직과 개인을 포함하여 AI 시스템 주기에서 능동적인 역할을 하는 사람	OECD Aritificial Intelligence in Society(2019)
책임감 있는 AI	• 공평하고 책임질 수 있는 기술의 결과를 의미하며 이를 통해 조직의 업무가 ISO에서 정의한 "직업적 책임*"에 따라 수행될 것으로 기대 가능 * AI 시스템 · 애플리케이션 · AI 기반 제품 또는 시스템을 설계 · 개발 · 배포하는 전문가가 개인 · 사회 · AI 미래에 영향을 미칠 수 있는 영향력을 인식하는 접근 방법	ISO/IEC TR 24368:2022
T.E.V.V	• 테스트, 평가, 검증, 유효성 검사(Test, Evaluation, Verification, and Validation)의 약어로 AI 시스템이나 컴포넌트를 검사하거나 수정하는 AI 행위자가 수행하는 작업	-

| 참고1 : 기존 SW의 위험과 AI 시스템 위험의 차이점 |

- **사회적 기술 |** 기존 소프트웨어와 비교했을 때 AI는 시간이 지남에 따라 예상치 못한 데이터로 훈련되어 시스템에 영향을 미칠 수 있는 '사회적 기술(Socio-tech)'로 사회 역학과 인간의 행동에 의해 영향받기 쉬움
  ※ AI는 기술적 요인뿐만 아니라 사회적 요인의 복잡한 상호작용에서 발생할 수 있으며, 온라인 챗봇 사용부터 구직, 대출 신청 등 사회 전 분야의 다양한 상황에서 사람들의 삶에 영향을 미칠 수 있음

- **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 문제가 AI 시스템의 신뢰성에 영향을 미칠 수 있음

- **데이터 의존성 |** AI 시스템은 학습데이터에 대한 종속성, 의존도가 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 사용할 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

- **규모와 복잡성 |** AI 시스템은 수십억 개 또는 수조 개의 결정 지점을 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

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

2. AI RMF의 주요내용

• 본 내용은 NIST가 공개한 「Artificial Intelligence Risk Management Framework(AI RMF1.0)」을 번역 · 요약

PART1 | 조직이 AI와 관련된 위험을 구성하는 방법 및 대상, 신뢰할 수 있는 AI 시스템의 특징을 설명
PART2 | 조직이 AI 시스템의 위험을 해결하는 데 도움이 되는 '거버넌스 · 매핑(위험식별) · 측정 · 관리' 등 4개 기능을 중심으로 AI 위험을 관리하고 신뢰할 수 있는 AI 시스템을 개발하기 위한 활동을 설명

PART 1 기본 정보

위험구성

① 위험, 영향 및 손해의 파악 및 해결

위험관리란 위험요소에 대해 조직을 지시하고 통제하기 위한 활동을 의미
(ISO 31000:2018)

○ AI RMF에서 정의하는 '위험'은 이벤트의 발생 확률과 그 결과의 규모 또는 정도에 대한 복합적인 측정값을 의미하며, AI 시스템의 영향이나 결과는 긍정적, 부정적 또는 양면적일 수 있어 이로 인해 기회 또는 위협이 초래될 수 있음(출처: ISO 31000:2018)

- 부정적 영향을 고려할 때 위험은 ▲이벤트나 상황이 발생할 경우 나타나는 부정적인 영향 또는 피해 규모, ▲발생 가능성을 의미하며, 개인 · 조직 · 생태계에서 발생 가능

○ 일반적인 위험관리 프로세스는 부정적인 영향을 다루나 AI RMF는 예상되는 부정적인 영향을 최소화하고 긍정적인 영향을 극대화할 수 있는 접근 방식을 제공하고자 함

- 위험관리를 통해 AI 개발자와 사용자는 모델과 시스템에 내재된 한계와 불확실성을 파악할 수 있어 전반적인 AI 시스템 성능 · 신뢰성을 향상시키고 및 AI 기술이 유익한 방식으로 사용될 가능성을 높일 수 있음

| 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 · 잡단/커뮤니티: 소수 인종, 민족 집단 차별 등 집단에 미치는 피해 · 사회: 민주적 참여 또는 교육적 접근성에 미치는 피해	· 조직의 비지니스 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

② 위험관리 과제

○ AI 위험관리를 위해서는 ▲위험 측정, ▲위험 허용 범위 설정, ▲위험 우선순위 지정, ▲조직적 통합 및 위험관리를 고려해 과제를 설정해야 함

- **(위험측정)** AI의 위험·오류를 정량적·정성적으로 측정하는 것은 어려우나 AI 위험을 측정할 수 없다고 하여 AI 시스템이 낮은 위험을 내포하고 있다는 의미는 아님

| 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 투명성·문서화 부족, AI 시스템에 내재된 불확실성으로 인해 발생할 수 있음
인간 기준선	• 의사결정 등 사람의 활동을 보강하거나 대체하기 위해 기준 설정

- **(위험 허용 범위)** 목표를 달성하기 위해 위험을 감수하려는 조직·AI 행위자의 준비로서 AI 시스템 소유자·조직·산업·커뮤니티 또는 정책 입안자가 설정한 정책 및 규범의 영향을 받으며 변경될 수 있음

- **(위험 우선순위)** 조직이 AI 시스템에 대해 가장 높게 판단한 위험을 기반으로 가장 시급한 위험의 우선순위와 위험관리 프로세스를 지정할 수 있음

  · AI 시스템이 허용할 수 없는 위험 수준*을 나타내는 경우 ⇒ 충분히 관리할 수 있을 때까지 해당 시스템의 개발 및 배포를 중단

  * 예: 심각한 악영향에 임박한 경우, 실제 심각한 피해가 발생한 경우, 치명적 위험 요소가 존재하는 경우

  · AI 시스템이 개인 정보 식별과 같은 민감한 또는 보호된 데이터로 구성된 대규모 데이터셋이거나 AI 시스템 결과가 인간에게 직접적·간접적으로 영향을 미치는 경우 ⇒ 높은 우선 순위

  · 전산 시스템과 상호 작용하도록 설계되며 민감하지 않은 데이터셋(예: 물리적 환경에서 수집된 데이터)에서 훈련되는 AI 시스템의 경우 ⇒ 비교적 낮은 우선 순위

  · 잔류위험을 문서화하기 위해 공급자는 AI 시스템 위험성을 충분히 고려하여 최종 사용자에게 시스템 상호 작용과 관련된 잠재적인 부정적인 영향을 알려야함

- **(조직적 통합 및 위험관리)** AI 위험관리는 AI 시스템을 개발하는 조직에 국한할 것이 아니라 사이버 보안 및 개인정보보호 등의 다른 중요한 위험 요소와 함께 전사적 위험관리 전략 및 프로세스에 통합되어야함
  ※ 다른 유형의 소프트웨어 개발 및 배포 시에 발생하는 위험과 중복되는 위험이 있을 수 있음

☑ 대상

○ AI 위험 및 잠재적인 긍정적 · 부정적 영향을 식별하고 관리하기 위해서는 AI 주기 전반에 걸친 광범위한 관점에서의 행위자가 필요

| AI 시스템의 주기 및 주요 차원 |

※ NIST는 OECD의 '인공지능 시스템 분류 및 프레임워크'(2022)를 참고해 AI의 주기 및 행위자를 설정했으며 테스트, 평가, 확인, 검증 프로세스의 중요성을 강조하고 일반화하기 위해 일부 수정

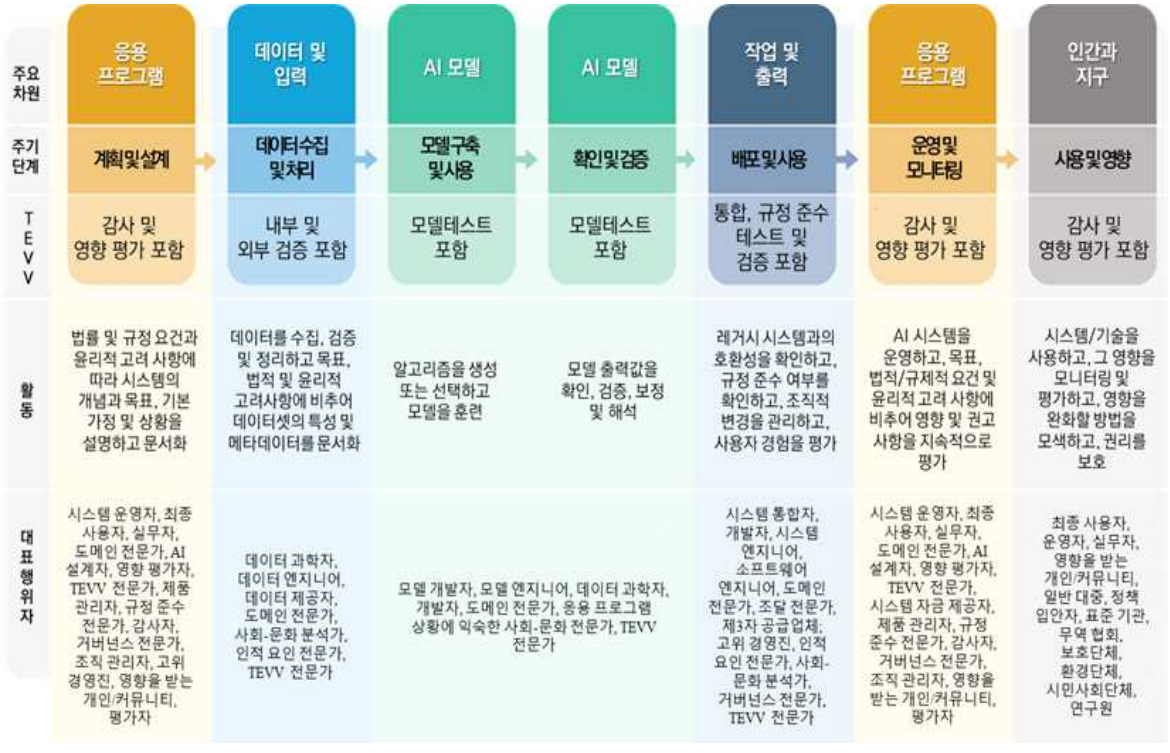
○ 내부 2개의 원은 AI 시스템의 주요 차원이며, 가장 외부의 원은 AI 주기를 나타냄. AI 주기에서 AI 위험 관리 활동을 주도하는 AI 행위자가 AI RMF의 주요 대상

| 참고2 : OECD 인공지능 시스템 분류 및 프레임워크의 5개 분류 차원 |

- **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

| AI 주기 전반에 걸친 AI 행위자 |



※ TEVV : 테스트(Test), 평가(Evaluation), 확인(Verification), 검증(Validation)

☑ 신뢰할 수 있는 AI 시스템의 특징

○ 신뢰할 수 있는 AI 시스템은 ▲타당하고 신뢰할 수 있으며, ▲안전하고, ▲보안이 철저하고, 탄력적이며, ▲책임을 질 수 있고, 투명하고, ▲설명 및 해석이 가능하고, ▲개인정보보호가 강화되고, ▲유해한 편향 관리를 통한 공정한 특성을 가짐

- 모든 특성은 시스템의 사회적 기술(Socio-tech) 속성을 가지나, 책임감 · 투명성은 외부 조건 및 AI 시스템 내부의 프로세스 · 활동과도 관련

- 신뢰성은 사회적 · 조직적 행동, AI 시스템에서 사용하는 데이터 셋, AI 모델 및 알고리즘의 선택과 이를 구축하는 사람들의 의사 결정, 시스템에서 인사이트를 제공하고 감독하는 사람과의 상호 작용과 밀접한 관련

※ AI 특성과 관련된 특정 지표 및 해당 지표에 대한 정확한 임계값은 인간의 판단을 통해 결정

- 신뢰할 수 있는 AI 시스템의 모든 특성은 서로 영향을 미치기 때문에 특성 하나 하나를 개별적으로 해결한다고 해서 AI 시스템의 신뢰성이 보장되는 것이 아님

※ 매우 안전하지만 불공정한 시스템, 정확하지만 불투명하고 해석할 수 없는 시스템, 부정확하지만 안전하고 개인정보 보호가 강화되고 투명한 시스템 등 특성 간의 균형이 이루어지지 않는 시스템은 모두 신뢰할 수 없음

| 신뢰할 수 있는 AI 시스템 특성 |

※ '유효성 및 신뢰성'은 신뢰할 수 있는 AI 시스템의 필수 조건이며, 이는 다른 특성의 기반
'책임감 및 투명성'은 다른 모든 특성과 관련이 있기 때문에 세로 상자로 표시

① **(유효성 및 신뢰성)** 배포된 AI 시스템의 유효성과 신뢰성은 시스템이 용도에 따라 작동하는지 확인하는 테스트 또는 모니터링을 통해 지속적으로 평가되며 유효성, 정확성, 견고성, 신뢰성을 측정하여 신뢰할 수 있는 AI 시스템을 구축할 수 있음

※ AI 위험관리 활동은 잠재적인 악영향을 최소화하는 것에 우선순위를 두어야하며 AI 시스템이 오류를 감지하거나 수정할 수 없는 경우 사람이 개입해야함

| '유효성 및 신뢰성' 관련 정의 |

정 의		출처
유효성 검사	객관적인 증거를 통해 특정 용도 또는 응용 프로그램에 대한 요구 사항이 충족되었음을 확인하는 것	ISO 9000:2015
신뢰성	주어진 조건 하에 주어진 시간 동안 실패 없이 필요에 따라 수행할 수 있는 능력	ISO/IEC TS 5723:2022
정확성	참값 또는 참이라고 인정되는 값에 대한 관찰, 계산 또는 추정의 근접성	ISO/IEC TS 5723:2022
견고성 (일반화 가능성)	다양한 상황에서 성능 수준을 유지하는 시스템 능력	ISO/IEC TS 5723:2022

② **(안전성)** AI 시스템은 정의된 조건에 따라 작동할 때 인간의 생명, 건강, 재산 또는 환경에 유해한 영향을 미치지 않아야 하며 AI 시스템의 안전성은 이하를 통해 개선 가능(출처 : ISO/IEC TS 5723:2022)

- 책임감 있는 설계, 개발 및 배포 기준
- 책임감 있는 시스템 사용에 대해 배포자에게 명확한 정보 제공
- 배포자 및 최종 사용자의 책임감 있는 의사 결정
- 사건의 실증적 증거를 기반으로 위험 설명 및 문서화

③ **(보안 및 탄력성)** AI 시스템 및 시스템이 배포된 생태계는 예상치 못한 이상 반응 또는 변화를 견딜 수 있는 경우, 내부/외부적 변화에도 불구하고 그 기능과 구조를 유지하며, 필요한 경우 안전하게 작동하면서 기능을 제한적으로 제공할 수 있는 탄력성이 있어야 함(출처: ISO/IEC TS 5723:2022)

※ 탄력성이 예상치 못한 이상 반응 이후 정상 기능으로 회복하는 능력이라면 보안은 탄력성과 함께 공격을 방지하고, 공격으로부터 보호하고, 공격에 대응하고, 그로부터 복구하는 프로토콜이 포함되어야 함

④ **(책임 및 투명성)** 책임감은 투명성*을 전제로 하며 부정확하거나 부정적인 영향을 초래하는 AI 시스템 결과를 시정하기 위해서 필요

* AI 시스템 및 그 결과에 대한 정보가 시스템과 상호작용하는 개인에게 제공되는 정도를 반영

- AI 주기 단계를 기반으로 AI 시스템과 상호 작용하거나 이를 사용하는 AI 행위자 또는 개인의 역할이나 정보에 맞게 조정된 정보에 대한 액세스를 제공하는 투명성은 더 높은 수준의 이해를 촉진하며 AI 시스템의 신뢰도를 높임
- 책임 및 투명성의 범위는 설계 의사결정, 훈련 데이터, 모델 훈련, 모델 구조, 의도한 사용 사례, 배포 시 또는 배포 후 최종 사용자의 의사 결정이 언제, 어떻게, 누구에 의해 이루어졌는지까지 다양
- 투명성을 위해서는 AI 시스템으로 인한 잠재적/실제적 악영향이 감지될 때 운영자 또는 사용자에게 이를 알리는 방법 등 인간-AI의 상호 작용을 고려해야 함
- AI 시스템에 대한 위험과 책임은 문화적, 법적, 부문별 및 사회적 관점에 따라 크게 달라지며 심각한 결과가 예상되는 경우* AI 개발자 및 배포자는 투명성 및 책임 기준을 비례적으로 사전에 미리 조정할 수 있어야 함

* 예: 생명 및 자유가 위협을 받는 경우

- AI 시스템 개발자는 해당 시스템이 의도한 대로 사용되는지 확인하기 위해 AI 배포자와 협력하여 다양한 유형의 투명성 도구를 테스트하는 것이 필요

⑤ **(설명 및 해석가능성)** AI 시스템 작동의 기본 메커니즘을 나타내는 '설명가능성'과 기능적 목적 관점에서 AI 시스템의 결과를 나타내는 '해석가능성'의 특성이 필요

※ 투명성은 시스템에서 "무슨 일이 발생했는지", 설명가능성은 "어떻게 결정이 내려졌는지", 해석가능성은 시스템이 결정을 내린 "이유와 그 의미 또는 상황"에 대한 질문에 답함

- 설명 가능성이 부족하여 발생하는 위험은 개인별 설명(예: 사용자의 역할, 지식 및 기술 수준에 따라 조정된 설명)과 함께 AI 시스템이 어떻게 작동되는지 설명함으로써 관리 가능

※ 설명 가능한 시스템은 보다 쉽게 디버깅 및 모니터링할 수 있으며 철저한 문서화, 감사 및 거버넌스에 적합

- 해석 가능성이 부족하여 발생하는 위험은 AI 시스템이 특정 예측이나 권고를 한 이유에 대해 설명함으로써 해결 가능

⑥ **(개인정보보호 강화)** AI 시스템을 설계, 개발 및 배포 시 익명성, 기밀성 및 제어 등 개인정보보호 가치 고려 필요

- 개인정보보호 관련 위험은 보안, 편향성, 투명성에 영향을 미칠 수 있으며, 안전 및 보안과 마찬가지로 특정 기술적 기능이 개인정보보호를 감소시킬 수 있음

- AI 시스템 설계 시 모델 결과값에 대한 비식별화 및 집계 등의 데이터 최소화 방법과 더불어 AI용 개인정보보호 강화 기술("PET", Privacy-enhancing technologies)은 개인정보보호 강화에 도움

⑦ **(공정성: 유해한 편향관리)** 3가지 범주의 편향(시스템적 편향, 통계적 편향, 인간 인지적 편향)을 고려하고 관리해야 함

- 시스템적 편향은 AI 데이터셋, AI 주기 전반에 걸친 조직적 규범, 수행 기준 및 절차, AI 시스템을 사용하는 광범위한 사회에 존재 가능
- 통계적 편향은 AI 데이터셋 및 알고리즘 프로세스에 존재할 수 있으며, 이는 대표성이 없는 샘플의 체계적 오류로 인해 종종 발생
- 인간 인지적 편향은 개인 또는 집단이 AI 시스템 정보를 인지하여 결정을 내리거나 누락된 정보를 채워 넣는 방법 또는 인간이 AI 시스템의 목적 및 기능에 대해 사고하는 방법과 관련이 있으며 AI의 설계, 구현, 운영 및 유지 관리를 포함하여 AI 주기 및 시스템 사용 전반에 걸친 의사 결정 프로세스에 편재

PART 2 핵심 및 프로파일

○ 신뢰할 수 있는 AI 시스템을 개발하기 위해 '거버넌스-매핑(위험식별)-측정-관리'의 4가지 핵심 기능을 구성하며 각 기능은 범주와 하위 범주로 구분됨

○ 위험관리는 AI 시스템 주기 전반에 걸쳐 수행되어야하며 핵심 기능은 외부조직의 AI 행위자 관점을 포함하여 다양하고 학제적 관점을 반영하는 방식으로 수행되어야 함

| 신뢰할 수 있는 AI 시스템 개발을 위한 4가지 핵심 기능 |

구 성	기 능
거버넌스	거버넌스는 3가지 다른 기능에 정보를 제공하고 제공받는 교차 기능으로 설계되었음. 조직 내 AI시스템을 설계, 개발, 배포 또는 획득하기 위한 위험관리 문화를 조성
관리	식별된 위험을 처리하고 시스템 오류 및 부정적 영향에 대한 가능성을 최소화하기 위해 거버넌스에서 설정된 문서 작성 기준, 매핑의 상황별 정보 및 측정의 경험적 정보를 활용. 관리 기능 후에는 위험 우선 순위를 지정하고 이를 지속적으로 모니터링 및 개선하기 위한 계획 수립
매핑 (위험 식별)	측정, 관리를 기초 작업으로 AI시스템의 위험, 광범위한 위험 유발 요인을 식별
측정	정량적, 정성적 또는 복합적 도구, 기법 및 방법을 채택하여 AI 위험 및 관련 영향을 분석하고 평가하고 벤치마킹하며 모니터링. 매핑에서 식별된 AI 위험과 관련한 지식을 활용해 관리를 위한 위험 모니터링

① **(거버넌스)** 거버넌스는 AI 시스템 수명 및 조직 계층 구조 전반에 걸쳐 AI 위험을 효과적으로 관리하기 위한 지속적이고 본질적인 요구 사항으로, 다른 핵심 기능을 가능하게 하며 조직 내에서 AI 시스템을 설계 · 개발 · 배포 · 획득하기 위한 위험관리 문화를 조성 가능하도록 해야함

| 거버넌스 기능 범주 및 하위범주|

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	**3.1** 주기 전반에 걸쳐 AI 위험을 매핑, 측정 및 관리하는 의사 결정을 내릴 때 다양한 부서로부터 정보를 얻는다.

범주(기능)	하위범주
위해 인력 다양성, 형평성, 포용성 및 접근성 프로세스의 우선 순위를 설정한다.	3.2 인간-AI 구성 및 AI 시스템 감독과 관련된 역할과 책임을 정의하고 차별화하기 위한 정책 및 절차를 마련한다.
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.

② **(매핑)** 매핑 기능을 수행하는 동안 수집된 정보는 부정적인 위험을 방지하고 프로세스(예: 모델 관리)에 대한 의사 결정은 물론 AI 솔루션의 적합성 또는 필요성에 대한 초기 의사 결정의 정보를 제공하며 매핑 기능에서의 결과물은 측정(MEASURE) 및 관리(MANAGE) 기능을 위한 토대를 형성함

매핑 기능 범주 및 하위범주

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.

범주(기능)	하위범주
	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.

③ **(측정)** 정량적, 정성적 또는 복합적 도구, 기법 및 방법을 채택하여 AI 위험 및 관련 영향을 분석, 평가, 모니터링하고 매핑(MAP) 기능에서 식별된 AI 위험과 관련한 지식을 사용하며 관리(MANAGE) 기능의 위험 모니터링 및 대응 활동에 필요한 정보를 제공

| 측정 기능 범주 및 하위범주|

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.

범주(기능)	하위범주
	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.

④ **(관리)** 식별된 위험을 처리하고 시스템 오류 및 부정적 영향에 대한 가능성을 최소화하기 위해 거버넌스에서 설정된 문서 작성 기준, 매핑의 상황별 정보 및 측정의 경험적 정보를 활용하며 관리 기능이 완료되면 위험 우선순위를 지정하고 이를 지속적으로 모니터링 및 개선하기 위한 계획이 수립됨

| 관리 기능 범주 및 하위범주|

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.

3. AI RMF의 로드맵

○ '신뢰할 수 있고 책임감 있는 AI 리소스 센터'(AIRC, Trustworthy and Responsible AI Resource Center) 출범('23.3월)해 AI RMF를 업데이트하고, 조직이 실행할 수 있도록 지원

○ 국제 표준을 따르고 관련 표준과의 연계를 지향*하며 신뢰성 특성 간의 균형을 맞추고 절충점을 파악하는 지침을 개발하고자 함

* 예: ISO/IEC 5338, ISO/IEC 38507, ISO/IEC 22989, ISO/IEC 24028, ISO/IEC DIS 42001, ISO/IEC NP 42005

○ AI 및 시스템의 신뢰성과 관련된 위험을 평가하기 위해 필요한 도구, 기준, 실험 환경 및 표준화된 방법론을 개발하기 위해 넓은 커뮤니티와 협력 예정

○ AI RMF 1.0 프로파일은 조직이 어떻게 실천해야하는지에 대한 사례이기 때문에 조직과 개인이 공동 또는 독립적으로 AI RMF 프로파일을 생성할 것을 권장

○ 프로그램 평가 전문가, AI RMF 사용자 커뮤니티 및 이해관계자들과 협력해 AI RMF의 효과를 측정하고 공유하는 방안을 수립 예정

○ 조직과 AI 행위자가 AI RMF를 실제로 어떻게 사용했는지에 대한 구체적인 사례를 발굴하고 연구할 예정

○ 개인 · 그룹 · 커뮤니티 · 사회에 대한 부정적인 영향/피해 가능성을 줄이기 위해 인간-AI 팀의 구성 연구, AI 설명가능성 및 해석가능성과 관련된 지침, 합리적인 위험 허용오차를 개발하는 방법에 대한 지침 연구 예정

○ AI 위험관리에 대한 다학제적, 사회기술적 접근 방식을 강화하기 위해 주제별 전문가 및 기타 이해관계자가 활용할 수 있는 자료, 지침 및 기타 리소스 제공 예정

| 참고3 : 미국의 AI 위험관리 정책 동향 |

연방정부 내 신뢰할 수 있는 인공지능의 활용을 촉진하기 위한 행정명령('20.12월)

· 행정절차 효율화 등을 위핸 연방행정기관이 AI를 이용하는 것을 허용하기 위한 행정명령
· 해당 AI의 설계, 취득, 개발 과정에서 미국 시민을 보호하기 위한 원칙으로 ▲합법성, ▲목적·성과중심, ▲신뢰성·효과성, ▲이해가능성 등을 제시

'2020 국가 AI 이니셔티브법' 제정('21.1월)

· 공공 및 민간 영역에서 신뢰할 수 있는 AI 시스템 개발 등을 위해 국가 AI 이니셔티브를 수립·이행하는 관련 활동을 명시하며 국가 이니셔티브실 설치. 국가 인공지능 위원회 설치, NIST의 인공지능 시스템에 대한 자발적 표준 개발 등을 포함

과학기술정책국(OSTP) 'AI 권리장전 청사진' 발표 ('22.10월)

· 미국 백악관은 인공지능(AI) 기술의 개발과 사용 과정에서 발생할 수 있는 부작용을 최소화하기 위한 'AI 권리장전(AI Bill of Rights) 청사진'을 발표
· ▲안전하고 효율적인 시스템 ▲알고리즘의 차별 방지 기능 ▲데이터 개인정보 보호 ▲사전 고지와 설명 ▲인적 대안 및 대체방안의 원칙으로 구성된 행정 조치

국립표준기술연구소(NIST) 'AI 위험관리 프레임워크' 발표('23.1월)

· 모든 분야와 규모의 기업 및 조직이 AI 위험을 해결할 수 있도록 유연하고 체계적이며 측정 가능한 프로세스를 제공하여 AI 기술의 이점을 극대화 하는 동시에 개인, 그룹, 지역사회, 조직 및 사회에 부정적인 영향을 미칠 가능성을 줄이도록 하기 위해 수립

국립과학재단(NSF) '국가 인공지능연구자원(NAIRR) 실행계획' 발표('23.1월)

· 혁신 촉진, 인재 다양성 증대, 역량 향상, 신뢰할 수 있는 AI 발전 등 4대 목표를 제시
· 시스템 안전 장치 구현을 위해 NIST의 5대 안전 프레임워크(안전한 프로젝트, 안전한 사람, 안전한 설정, 안전한 데이터, 안전한 결과물) 준수, 국가 인공지능 연구자원 사이버 인프라를 "NAIRR-Open'(개방) 영역과 "NAIRR-Secure"(보안) 영역으로 구분하여 설계

연방 정부를 통한 인종 형평성 증진 및 소외된 지역사회 지원에 관한 행정 명령('23.2월)

· 바이든 대통령은 연방 기관이 AI를 포함한 신기술의 설계와 사용에서 편견을 근절하고 알고리즘 차별로부터 대중을 보호하도록 지시하는 행정 명령에 서명

자동화 시스템의 차별 및 편견에 대한 집행 노력에 관한 공동 성명서('23.4월)

· 연방거래위원회, 소비자금융보호국, 평등고용기회위원회, 법무부 민권국은 공동 성명을 발표하여 기존의 법적 권한을 활용하여 AI 관련 피해로부터 미국 국민을 보호하겠다는 공동의 약속을 강조

과학기술정책국(OSTP), '국가 인공지능 전략' 수립을 위한 의견 수렴('23.5)

· 미국 연방정부는 최신 AI 기술발전에 따른 위험기회 및 전 세계적 변화에 대비하기 위한 범사회적 대응 방법으로 '국가 AI 전략' 활용 예정

· 과학기술정책국(OSTP)는 AI 연관된 정부의 우선순위 및 향후 조치방안에 과한 의견을 공개적으로 수렴(RFI) 중이며 이 중 '국민 권리, 안전 및 국가 안보 보호 보장 등 위험관리'에 관한 정보제공요청이 가장 많음

백악관, '미국인의 권리와 안전을 보호하는 책임감있는 AI 혁신을 촉진하기 위한 조치' 발표('23.5월)

① 책임감 있는 미국 AI 연구 및 개발(R&D)을 위한 새로운 투자

· 연방정부 기관인 국립과학재단(NSF)에서 1억4000만 달러(약 1조8500억 원)를 출자해 7개의 국립 AI 연구기관을 신설(본 투자로 미국 전역에 총 25개의 연구소 설립)

· 연구기관은 고등 교육 기관, 연방 기관, 산업계 및 기타 기관 간의 협력을 촉진하여 윤리적이고 신뢰할 수 있으며 책임감 있고 공익에 기여하는 혁신적인 AI 발전을 추구

· 미국의 AI R&D 인프라를 강화하고 다양한 AI 인력의 개발을 지원

② 생성 AI 시스템에 대한 공개 평가

· 7개 하이테크 기업*은 자사의 AI 시스템에 사용한 학습데이터를 대중에 공개하고, 행정부가 도입하는 규제에 부합하는지를 확인하기 위해 평가를 받기로 합의

· 수천 여명의 커뮤니티 파트너와 AI 전문가가 AI 모델을 철저히 평가하여 해당 모델이 바이든 행정부의 AI 권리장전 및 AI 위험관리 프레임워크에 명시된 원칙과 관행에 어떻게 부합하는지 확인 예정

* 구글, 마이크로소프트, 오픈AI, 스태빌리티 AI, 엔비디아, 허깅 페이스, 앤트로픽

· 평가는 민간 전문기업인 '스케일AI'가 주관하며 라스베가스에서 열리는 세계적 해커 행사 '데프콘 2023'('23.8.10~13)의 부대행사인 'AI빌리지'*에서 이뤄질 전망

* 데이터과학자와 해커들이 참여해 만든 비영리 커뮤니티

③ 미국 정부의 AI 시스템 사용에 관한 정책 지침 초안 발표 예정

· 미국 관리예산처(OMB)는 미국 정부의 AI 시스템 사용에 관한 정책 지침 초안(draft policy guidance on the use of AI systems by the U.S. government for public comment)을 공개하고 대중의 의견을 수렴할 예정이라고 발표

· 미국 국민의 권리와 안전을 보호하는 데 중점을 둔 AI 시스템의 개발, 조달 및 사용을 보장하기 위해 연방 부처와 기관이 따라야 할 구체적인 정책을 수립 예정

4. 주요 시사점

1 AI 기술의 특징에 맞는 적극적인 위험관리 정책 추진

○ 미국은 디지털 신기술의 연구·개발·시장창출을 민간에서 주도하지만 AI의 경우 정부가 AI RMF와 같은 구체적인 가이드라인과 후속 조치들을 발표하며 적극적으로 위험관리 정책을 추진

- 그간 과잉규제 지양과 위험 기반 사후규제 기조 하에 AI 신뢰성 확보를 위한 원칙, 가이드라인을 발표* 했으나 AI RMF 수립과 함께 위험을 사전 방지해야 한다는 의견과 정책**을 제시

* 연방정부 규제 가이드라인(백악관, '20.1월)

** 백악관은 '우리 시대의 가장 강력한 기술 중 하나인 AI가 제시하는 기회를 포착하기 위해서 먼저 위험을 완화할 필요가 있기에 기업이 AI 제품을 배포·공개 전에 안전성을 확인해야 할 근본적 책임'이 있다고 표명('23.5월)

○ AI는 기존 SW와 다른 속성의 위험*을 내포하고 있어 이에 맞춰 지속적이고 유연하게 업데이트할 수 있는 새로운 위험관리의 방식의 정책적 가이드의 필요성 제시

* 데이터 품질 문제, 데이터에 대한 의존성, 성능 예측의 불확실성, 규모와 복잡성, 개인정보 이슈 등

- AI RMF에서는 AI를 'Socio-tech'(사회적 기술)라고 칭하며 기술적 요인뿐만 아니라 사회적 요인의 복잡한 상호작용을 통해 작동하며, 사회 전 분야의 다양한 상황에서 사람들의 삶에 영향을 미치는 기술로 강조

'Socio-tech(사회적 기술)'로서의 AI

• AI 시스템은 사회적 환경과 인간 행동의 영향을 받아 작동하는 사회적 기술로써 시간이 지남에 따라 사회적 영향을 받은 예상치 못한 데이터로 훈련되어 시스템이 작동하기도 하고, 시스템의 결과가 사회 전 분야의 다양한 상황에서 사람들의 삶에 영향을 끼칠 수 있는 상호작용을 하는 기술

출처 : AI RMF('23.1), 저자 재정의

2 신뢰할 수 있는 AI 시스템 구축을 위한 포괄적이고 유연한 가이드 제시

○ AI 시스템은 전 과정*의 전 행위자**가 AI 시스템의 신뢰성 확보를 위한 위험관리를 고려해야 완전히 작동할 수 있기 때문에 이를 포괄하는 위험관리 프로파일을 제공했으며 모범 적용 사례도 발굴 예정

* 계획 및 설계, 데이터 수집 및 처리, 모델 구축 및 사용, 확인 및 검증, 배포 및 사용, 운영 및 모니터링, 사용 및 영향평가

** 고위 경영진, 시스템 개발자, 운영자, 최종 사용자 등

○ 가이드는 AI 활성화를 저해하는 규범적 성격이 아니라 조직의 자체적인 정책, 규범 등과 결합하여 사용될 수 있도록 유연성을 가질 수 있어야 함

3 AI RMF 수립부터 평가까지, 민관 협업으로 실효성 확보

○ AI RMF1.0은 민관 협업으로 지속적인 의견 수렴과 공개 과정을 거쳐 수립되었으며 향후 평가 또한 민간 전문기업이 위탁하여 다양한 전문가들이 참여하는 커뮤니티를 통해 이루어질 것으로 예정된바 민관 협업을 중시

- 7개의 하이테크 기업(구글, 마이크로소프트, 오픈AI, 스태빌리티 AI, 엔비디아, 허깅 페이스, 앤트로픽)은 백악관과의 협의를 통해 자발적으로 AI RMF1.0을 포함해 정부가 설정한 AI 위험관리 방향에 부합하는지 평가받을 예정(백악관, '23.5월 발표*)

* THE WHITE HOUSE('23.5월), ACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety

※ AI RMF1.0의 본문 및 로드맵에서도 미래 NIST 활동의 일환으로 AI 커뮤니티와 함께 AI RMF의 효율성(AI 시스템의 신뢰도 향상을 측정하는 방법 포함)을 평가 예정임을 밝힘

④ 신뢰할 수 있는 AI를 위한 지원 체계 구성

○ NIST는 AI RMF를 지속적으로 갱신하고, 다양한 조직이 실행할 수 있도록 지원하는 '신뢰할 수 있고 책임감 있는 AI 리소스 센터'(AIRC, Trustworthy and Responsible AI Resource Center) 출범('23.3월)

- AI 관련 국제 표준, 연방 지침, 신뢰할 수 있는 AI 시스템 구축을 위한 공공 · 민간의 발행 자료의 저장소 역할을 하고 AI RMF를 갱신하기 위한 의견 수렴, 평가 등을 추진

〈 참 고 자 료 〉

[1] NIST(2023.1.26.). NIST Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence, https://www.nist.gov/news-events/news/2023/01/nist-risk-management-framework-aims-improve-trustworthiness-artificial

[2] NIST Trustworthy & Responsible AI Resource Center, https://airc.nist.gov/home

[3] CIO(2022.4.14.). NIST, AI 위험관리 프레임워크 마련, https://www.ciokorea.com/tags/2760/ai/232439#csidxbf879700c8f66bb8061ae33996a85fd

[4] 우상근(2022.4), OECD 인공지능 시스템 분류 프레임워크 분석 및 시사점, AI REPORT 2022-1

[5] THE WHITE HOUSE(OSTP, 2022.10.31), Blueprint for an AI Bill of Rights, https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[6] THE WHITE HOUSE(2023.5.4). FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety, https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/

[7] THE WHITE HOUSE(2023.2.16). Executive Order on Further Advancing Racial Equity and Support for Underserved Communities Through The Federal Government. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/02/16/executive-order-on-further-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/

[8] U.S. Equal Employment Opportunity Commission(2023.4). joint statement on enforcement efforts against discrimination and bias in automated systems. https://www.eeoc.gov/joint-statement-enforcement-efforts-against-discrimination-and-bias-automated-systems

[9] 과학기술인재정책플랫폼(2023.2.9.). NSF 국가 인공지능 연구 자원(NAIRR) 태스크포스 최종보고서 발표. https://hrstpolicy.re.kr/kistep/kr/board/BoardDetail.html?board_seq=53024&board_class=BOARD03&rootId=2006000&menuId=2006101

[10] 테크데일리(2023.5.5.). 美 정부, 'AI 위험' 대책 발표. https://www.techdaily.co.kr/news/articleView.html?idxno=22131

[11] ZDNET(2023.5.6.). 구글·MS·오픈AI 등 AI제품 알고리즘 대중에 공개된다. https://zdnet.co.kr/view/?no=20230506152252

[12] 법률신문(2023.5.11.). 미국의 AI 규제 동향 및 시사점. https://www.lawtimes.co.kr/news/187480

[13] Nextgov(2023.3.30.). NIST Debuts Trustworthy and Responsible AI Resource Center. https://www.nextgov.com/emerging-tech/2023/03/nist-debuts-trustworthy-and-responsible-ai-resource-center/384618/

THE AI REPORT 2023

NIA 한국지능정보사회진흥원 AI·미래전략센터