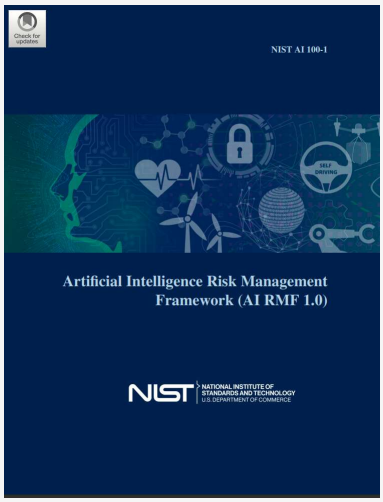
AI 리스크관리의 구조와 그 형식 (AI RMF 1.0)

K-Risk 발간편집 위원회



목차

2023년(봄)
2023년(여름)
2023년(가을)

K-Risk

※ 본 기사는 좌측 문헌의 단순 번역기사로서 K-Risk의 견해를 반영하는 것은 아니다.

※ 상기 이미지를 클릭하면 원문 다운로드가 가능합니다.

2부: 핵심 및 프로필

5. AI RMF 핵심

AI RMF 핵심(AI Risk Management Framework Core)는 AI 리스크를 관리하고 신뢰할 수 있는 AI 시스템을 책임감 있게 개발하기 위한 대화, 이해, 활동을 가능하게 하는 결과물과 조치를 제공한다. 그림 5에서 볼 수 있듯이 핵심은 GOVERN, MAP, MEASURE 및 MANAGE의 네 가지 기능으로 구성된다. 각 상위 레벨 기능은 카테고리와 하위 카테고리로 분류된다. 카테고리와 하위 카테고리는 구체적 조치와 그 결과로 세분화된다. 작업은 특별한 체크리스트가 없으며 반드시 순서가 지정된 단계가 아니다.

그림. 5 AI RMF의 핵심기능

그림 5. 기능은 AI 리스크를 GOVERN, MAP, MEASURE 및 MANAGE(통제, 매핑, 측정 및 관리)하기 위해 최고 수준에서 AI 리스크관리 활동을 조직한다. 거버넌스는 다른 세 가지 기능 전반에 걸쳐 정보를 제공하고 영향을 미치는 교차 기능으로 설계된다.

리스크관리는 지속적이고 시의적절하며 AI 시스템 생애주기 전반에 걸쳐 수행되어야 한다. AI RMF 핵심 기능은 잠재적으로 조직 외부의 AI 행위자의 견해를 포함하여 다양하고 다학문적인 관점을 반영하는 방식으로 수행되어야 한다. 다양한 팀을 구성하면 설계, 개발, 배포 또는 평가 중인 기술의 목적과 기능에 대한 아이디어와 가정을 보다 공개적으로 공유하는 데 도움이 된다. 이를 통해 문제를 표면화하고 기존 및 긴급 리스크를 식별할 수 있다.

AI RMF의 온라인 동반 자원인 NIST AI RMF 플레이북은 조직이 AI RMF를 탐색하고 자체 상황에 적용할 수 있는 제안된 전술적 조치를 통해 결과를 달성하는 데 도움이 된다. AI RMF와 마찬가지로 플레이북은 자발적이며 조직은 필요와 관심에 따라 제안을 활용할 수 있다. 플레이북 사용자는 추천 자료에서 자신이 사용할 맞춤형 지침을 선택하고 더 넓은 커뮤니티와 공유하기 위한 제안을 제공할 수 있다. AI RMF와 함께 플레이북은 신뢰할 수 있고 책임 있는 AI 리소스 센터의 일부이다.

프레임워크 사용자는 자원과 역량을 기반으로 AI 리스크를 관리하기 위한 요구사항에 가장 적합한 기능을 적용할 수 있다. 일부 조직은 카테고리 및 하위 카테고리 중에서 선택할 수 있다. 다른 조직들은 모든 카테고리와 하위 카테고리를 선택하고 적용할 수 있다. 거버넌스 구조가 확립되어 있다고 가정하면 프레임워크 사용자가 가치를 추가하는 기능은 AI 생애주기 전반에 걸쳐 어떤 순서로든 수행될 수 있다. GOVERN에서 결과를 설정한 후 대부분의 AI RMF 사용자는 MAP 기능으로 시작하여 계속해서 MEASURE(측정) 또는 MANAGE(관리)를 수행한다. 그러나 사용자는 기능을 통합하므로 필요에 따라 기능 간 상호 참조를 통해 프로세스가 반복되어야 한다. 마찬가지로, 여러 기능에 적용되거나 특정 하위 카테고리 결정 이전에 논리적으로 발생해야 하는 요소가 포함된 카테고리 및 하위 카테고리가 있다.

5.1 GOVERN

GOVERN 기능:

- AI 시스템을 설계, 개발, 배포, 평가 또는 획득하는 조직 내에서 리스크관리 문화를 조성하고 구현한다.
- 시스템이 사용자와 사회 전반에 걸쳐 제기할 수 있는 리스크를 예측, 식별 및 관리하는 프로세스, 문서 및 조직 계획과 이러한 결과를 달성하기 위한 절차를 간략하게 설명한다.
- 잠재적 영향을 평가하는 프로세스를 통합한다.
- AI 리스크관리 기능이 조직의 원칙, 정책 및 전략적 우선순위와 일치할 수 있는 구조를 제공한다.
- AI 시스템 설계 및 개발의 기술적 측면을 조직의 가치 및 원칙에 연결하고 해당 시스템의 획득, 교육, 배포 및 모니터링과 관련된 개인의 조직 실무 및 역량을 활성화한다. 그리고
- 타사 소프트웨어나 하드웨어 시스템 및 데이터 사용과 관련된 법적 문제 및 기타 문제를 포함하여 전체 제품 생애주기 및 관련 프로세스를 다룬다.

GOVERN은 AI 리스크관리 전반에 걸쳐 투입되어 프로세스의 다른 기능을 활성화하는 교차 기능이다. GOVERN의 측면, 특히 규정 준수 또는 평가와 관련된 측면은 각각의 다른 기능에 통합되어야 한다. 거버넌스에 대한 관심은 AI 시스템의 생애와 조직의 계층 구조에 걸쳐 효과적인 AI 리스크관리를 위한 지속적이고 본질적인 요구사항이다.

강력한 거버넌스는 내부 실무와 규범을 추진하고 강화하여 조직의 리스크 문화를 향상시킬 수 있다. 관리 당국은 조직의 임무, 목표, 가치, 문화 및 리스크 허용 범위를 지시하는 중요한 정책을 결정한다. 상위 리더십은 조직 내 리스크를 관리하고 이를 통해 조직 문화를 설정한다. 경영진은 AI 리스크관리의 기술적 측면을 정책 및 운영에 맞춰 조정한다. 문서화는 투명성을 높이고, 인적 검토 프로세스를 개선하며, AI 시스템 팀의 책임성을 강화한다.

GOVERN 기능에 설명된 구조, 시스템, 프로세스 및 팀을 배치한 후 조직은 리스크 이해 및 관리에 초점을 맞춘 목적 중심 문화의 이점을 누릴 수 있다. 시간이 지남에 따라 AI 행위자의 지식, 문화, 요구 또는 기대가 발전함에 따라 GOVERN 기능을 계속 실행하는 것은 프레임워크 사용자의 의무이다.

AI 리스크관리와 관련된 실무는 NIST AI RMF 플레이북에 설명되어 있다. 표 1에는 GOVERN 함수의 카테고리와 하위 카테고리가 나열되어 있다.

표 1 : GOVERN 기능을 위한 카테고리 및 하위 카테고리

카테고리	하위 카테고리
GOVERN 1: AI 리스크의 매핑, 측정 및 관리와 관련된 조직 전반의 정책, 프로세스, 절차 및 실무가 투명하고 효과적으로 구현되어 있다.	**GOVERN 1.1:** AI와 관련된 법률 및 규제 요구사항을 이해하고 관리하며 문서화한다. **GOVERN 1.2:** 신뢰할 수 있는 AI의 특성은 조직의 정책, 프로세스, 절차 및 실무에 통합된다. **GOVERN 1.3:** 조직의 리스크 허용 범위를 기반으로 필요한 리스크관리 활동 수준을 결정하기 위한 프로세스, 절차 및 실무가 마련되어 있다. **GOVERN 1.4:** 리스크관리 프로세스와 그 결과는 조직의 리스크 우선순위에 따라 투명한 정책, 절차 및 기타 통제를 통해 설정된다. **GOVERN 1.5:** 리스크관리 프로세스와 그 결과에 대한 지속적인 모니터링과 정기 검토가 계획되고 정기 검토 빈도 결정을 포함하여 조직의 역할과 책임이 명확하게 정의된다. **GOVERN 1.6:** AI 시스템의 목록을 작성하는 메커니즘이 마련되어 있으며 조직의 리스크 우선순위에 따라 자원이 배정된다. **GOVERN 1.7:** 리스크를 증가시키거나 조직의 신뢰성을 저하시키지 않는 방식으로 AI 시스템을 안전하게 폐기하고 단계적으로 폐지하기 위한 프로세스와 절차가 마련되어 있다.
GOVERN 2: 적절한 팀과 개인이 AI 리스크를 매핑, 측정 및 관리하기 위한 권한을 부여받고 책임을 지며 교육을 받을 수 있도록 책임 구조가 마련되어 있다.	**GOVERN 2.1:** AI 리스크 매핑, 측정 및 관리와 관련된 역할과 책임, 커뮤니케이션 라인이 문서화되어 조직 전체의 개인과 팀에 명확하게 전달된다. **GOVERN 2.2:** 조직의 직원과 파트너는 관련 정책, 절차 및 계약에 따라 직무와 책임을 수행할 수 있도록 AI 리스크관리 교육을 받는다. **GOVERN 2.3:** 조직의 경영진은 AI 시스템 개발 및 배포와 관련된 리스크에 대해 결정한다.

표 1 : GOVERN 기능을 위한 카테고리 및 하위 카테고리(계속)

카테고리	하위 카테고리
GOVERN 3: 생애주기 전반에 걸쳐 AI 리스크를 매핑, 측정 및 관리하는데 있어 인력 다양성, 형평성, 포용성 및 접근성 프로세스가 우선적으로 고려된다.	**GOVERN 3.1:** 생애주기 전반에 걸쳐 AI 리스크를 매핑, 측정 및 관리하는 것과 관련된 의사결정은 다양한 팀의 정보를 받는다. (예: 인구 통계, 분야, 경험, 전문 지식 및 배경의 다양성) **GOVERN 3.2:** 인간-AI 구성 및 AI 시스템 감독에 대한 역할과 책임을 정의하고 차별화하기 위한 정책과 절차가 마련되어 있다.
GOVERN 4: 조직은 AI 리스크를 고려하고 전달하는 문화에 전념한다.	**GOVERN 4.1:** 잠재적인 부정적 영향을 최소화하기 위해 AI 시스템의 설계, 개발, 배포 및 사용에 있어 비판적 사고와 안전 우선 사고방식을 육성하기 위한 조직 정책 및 실무가 마련되어 있다. **GOVERN 4.2:** 조직은 설계, 개발, 배포, 평가 및 사용하는 AI 기술의 리스크와 잠재적 영향을 문서화하고 그 영향에 대해 보다 광범위하게 소통한다. **GOVERN 4.3:** AI 테스트, 사고 식별 및 정보 공유를 지원하는 조직 실무가 마련되어 있다.
GOVERN 5: 관련 AI 행위자와의 참여 프로세스가 마련되어 있다.	**GOVERN 5.1:** AI 리스크와 관련된 개인 및 사회적 잠재적 영향에 관해 AI 시스템을 개발하거나 배포한 팀 외부의 피드백을 수집, 고려, 우선순위 지정 및 통합하기 위한 조직 정책 및 실무가 마련되어 있다. **GOVERN 5.2:** AI 시스템을 개발하거나 배포한 팀이 관련 AI 행위자로부터 판정된 피드백을 시스템 설계 및 구현에 정기적으로 통합할 수 있도록 하는 메커니즘이 확립되었다.
GOVERN 6: 제3자 소프트웨어와 데이터, 기타 공급망 문제로 인해 발생하는 AI 리스크와 이점을 해결하기 위한 정책과 절차가 마련되어 있다.	**GOVERN 6.1:** 제3자의 지적재산권이나 기타 권리 침해 리스크를 포함하여 제3자 단체와 관련된 AI 리스크를 해결하는 정책 및 절차가 마련되어 있다. **GOVERN 6.2:** 리스크가 높은 것으로 간주되는 제3자 데이터 또는 AI 시스템의 오류나 사고를 처리하기 위한 비상 프로세스가 마련되어 있다.

5.2 MAP

MAP 기능은 AI 시스템과 관련된 리스크를 프레임화하기 위한 맥락를 설정한다. AI 생애주기는 다양한 행위자가 참여하는 많은 상호의존적 활동으로 구성된다(그림 3 참조). 실제로 프로세스의 한 부분을 담당하는 AI 행위자는 다른 부분 및 관련 맥락에 대한 완전한 가시성이나 제어권을 갖지 못하는 경우가 많다. 이러한 활동과 관련 AI 행위자 간의 상호 의존성으로 인해 AI 시스템 영향을 안정적으로 예측하기가 어려울 수 있다. 예를 들어, AI 시스템의 목적과 목표를 식별하는 초기 결정은 시스템의 동작과 기능을 변경할 수 있으며, 배포 설정(예: 최종 사용자 또는 영향을 받는 개인)의 역학은 AI 시스템 결정의 영향을 형성할 수 있다. 결과적으로, AI 생애주기의 한 차원 내에서 최선의 의도는 다른 이후 활동의 결정 및 조건과의 상호작용을 통해 훼손될 수 있다.

이러한 복잡성과 다양한 가시성 수준으로 인해 리스크관리 실무에 불확실성이 발생할 수 있다. 부정적인 리스크의 잠재적 원인을 예측, 평가 및 해결하면 이러한 불확실성을 완화하고 의사결정 프로세스의 무결성을 향상시킬 수 있다.

MAP 기능을 수행하면서 수집된 정보는 부정적 리스크 예방을 가능하게 하고 모델 관리 등 프로세스에 대한 결정은 물론 AI 솔루션의 적합성이나 필요성에 대한 초기 결정을 알려준다. MAP 기능의 결과는 MEASURE 및 MANAGE 기능의 기반이다. 상황에 맞는 지식과 식별된 상황 내 리스크에 대한 인식이 없으면 리스크관리를 수행하기가 어렵다. MAP 기능은 리스크와 광범위한 기여 요인을 식별하는 조직의 능력을 향상시키기 위한 것이다.

이 기능의 구현은 다양한 내부 팀의 관점을 통합하고 AI 시스템을 개발하거나 배포한 팀 외부 사람들과의 참여를 통해 향상된다. 외부 협력자, 최종 사용자, 잠재적으로 영향을 받는 커뮤니티 등과의 참여는 특정 AI 시스템의 리스크 수준, 내부 팀 구성 및 조직 정책에 따라 달라질 수 있다. 이러한 광범위한 관점을 수집하면 조직은 다음을 통해 부정적인 리스크를 사전에 예방하고 보다 신뢰할 수 있는 AI 시스템을 개발하는 데 도움이 될 수 있다.

- 상황을 이해하는 능력을 향상시킨다.
- 사용 상황에 대한 가정을 확인한다.
- 시스템이 의도한 상황 내에서 또는 외부에서 작동하지 않는 경우를 인식할 수 있다.
- 기존 AI 시스템의 긍정적이고 유익한 용도를 구별해낸다.
- AI 및 ML 프로세스의 한계에 대한 이해를 향상시킨다.
- 부정적인 영향을 초래할 수 있는 실제 응용 프로그램의 제약 조건을 식별한다.
- AI 시스템의 의도된 사용과 관련하여 알려지고 예측 가능한 부정적인 영향을 식별한다.
- 의도된 용도를 넘어서 AI 시스템을 사용할 때 발생할 수 있는 리스크를 예상한다.

MAP 기능을 완료한 후 프레임워크 사용자는 AI 시스템을 설계, 개발 또는 배포할지 그 여부에 대한 초기 진행/중단 결정을 알리기 위해 AI 시스템 영향에 대해 충분히 알아야 한다. 진행하기로 결정한 경우 조직은 AI 리스크관리를 지원하기 위해 GOVERN 기능에 마련된 정책 및 절차와 함께 측정 및 관리 기능을 활용해야 한다. 시간이 지남에 따라 맥락, 기능, 리스크, 이점 및 잠재적 영향이 발전함에 따라 MAP 기능을 AI 시스템에 계속 적용하는 것은 프레임워크 사용자의 의무이다.

AI 리스크 매핑과 관련된 실무는 NIST AI RMF 플레이북에 설명되어 있다. 표 2에는 MAP 기능의 카테고리와 하위 카테고리가 나열되어 있다.

표 2 : MAP 기능을 위한 카테고리 및 하위 카테고리

카테고리	하위 카테고리
MAP 1: 맥락 확립 및 이해	**MAP 1.1:** AI 시스템 배포 의도의 목적, 잠재적으로 유익한 용도, 상황별 법률, 규범 및 기대, 향후 설정을 이해하고 문서화한다. 다음의 사항들을 고려해야 한다. 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 생애주기 전반에 걸쳐 AI 시스템 목적, 사용 및 리스크에 대한 가정 및 관련 제한사항 관련 TEVV 및 시스템 측정항목. **MAP 1.2:** 상황 설정을 위한 다학제간 AI 행위자, 역량, 기술 및 역량은 인구통계학적 다양성, 광범위한 영역 및 사용자 경험 전문 지식을 반영하며 이들의 참여가 문서화된다. 다학제간 협력 기회가 우선시된다. **MAP 1.3:** AI 기술에 대한 조직의 사명과 관련 목표를 이해하고 문서화한다. **MAP 1.4:** 비즈니스 가치 또는 비즈니스 사용 맥락이 명확하게 정의되었거나 기존 AI 시스템을 평가하는 경우 재평가된다. **MAP 1.5:** 조직의 리스크 허용 범위가 결정되고 문서화된다. **MAP 1.6:** 시스템 요구사항(예: "시스템은 사용자의 개인 정보를 존중해야 한다")은 관련 AI 행위자로부터 도출되고 이해된다. 설계 결정은 AI 리스크를 해결하기 위해 사회 기술적 영향을 고려한다.
MAP 2: AI 시스템 분류	**MAP 2.1:** AI 시스템이 지원할 작업을 구현하는 데 사용되는 특정 작업과 방법이 정의된다(예: 분류자, 생성 모델, 추천자). **MAP 2.2:** AI 시스템의 지식 한계와 인간이 시스템 출력을 활용하고 감독하는 방법에 대한 정보가 문서화되어 있다. 문서는 관련 AI

표 2 : MAP 기능을 위한 카테고리 및 하위 카테고리(계속)

카테고리	하위 카테고리
	행위자가 결정을 내리고 후속 조치를 취할 때 도움이 되는 충분한 정보를 제공한다. MAP 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 문서화된다.
MAP 3: 적절한 벤치마크와 비교를 통하여 AI 기능, 목표 사용법, 목표, 예상 이점 및 비용 이해	MAP 3.1: 의도된 AI 시스템 기능 및 성능의 잠재적 이점을 검사하고 문서화한다. MAP 3.2: 예상되거나 실현된 AI 오류나 시스템 기능 및 신뢰성(조직의 리스크 허용 범위와 관련)으로 인해 발생하는 비금전적 비용을 포함한 잠재적 비용을 검사하고 문서화한다. MAP 3.3: 시스템 기능, 확립된 맥락 및 AI 시스템 분류를 기반으로 대상 애플리케이션 범위가 지정되고 문서화된다. MAP 3.4: AI 시스템 성능 및 신뢰성, 관련 기술 표준 및 인증에 대한 운영자 및 실무자의 숙련도를 위한 프로세스가 정의, 평가 및 문서화된다. MAP 3.5: 인력 감독을 위한 프로세스는 GOVERN 기능의 조직 정책에 따라 정의, 평가 및 문서화된다.
MAP 4: 타사 소프트웨어 및 데이터를 포함하여 AI 시스템의 모든 구성 요소에 대한 리스크와 이점 매핑	MAP 4.1: 제3자의 지적재산권 또는 기타 권리 침해 리스크와 마찬가지로 AI 기술과 제3자 데이터 또는 소프트웨어의 사용을 포함하여 해당 구성 요소의 법적 리스크를 매핑하기 위한 접근 방식을 확립하여 이를 준수하고 문서화한다. MAP 4.2: 타사 AI 기술을 포함한 AI 시스템 구성 요소에 대한 내부 리스크 제어가 식별되고 문서화된다.
MAP 5: 개인, 그룹, 지역사회, 조직 및 사회에 미치는 영향 특성화	MAP 5.1: 예상되는 사용, 유사한 맥락에서 AI 시스템의 과거사용, 공공 사건보고, AI 시스템을 개발하거나 배포한 팀 외부 사람들의 피드백을 기반으로 식별된 각 영향(잠재적으로 유익하거나 유해할 수 있음)의 가능성 및 규모 또는 기타 데이터가 식별되고 문서화된다. MAP 5.2: 관련 AI 행위자와의 정기적인 참여를 지원하고 긍정적, 부정적, 예상치 못한 영향에 대한 피드백을 통합하기 위한 실무와 인력이 마련되어 문서화된다.

5.3 MEASURE

MEASURE 기능은 정량적, 정성적 또는 혼합 방법 도구, 기술 및 방법론을 사용하여 AI 리스크 및 관련 영향을 분석, 평가, 벤치마킹 및 모니터링한다. MAP 기능에서 식별된 AI 리스크와 관련된 지식을 활용하고 MANAGE 기능에 알려준다. AI 시스템은 배포 전과 작동 중에 정기적으로 테스트해야 한다. AI 리스크 측정에는 시스템 기능 및 신뢰성 측면의 문서화가 포함된다.

AI 리스크 측정에는 신뢰할 수 있는 특성, 사회적 영향 및 인간-AI 구성에 대한 지표 추적이 포함된다. MEASURE 기능에서 개발되거나 채택된 프로세스에는 관련 불확실성 측정, 성능 벤치마크 비교, 공식 보고 및 결과 문서화와 관련된 엄격한 소프트웨어 테스트 및 성능 평가 방법론이 포함되어야 한다. 독립 검토 프로세스는 테스트의 효율성을 향상시키고 내부 편견과 잠재적인 이해 상충을 완화할 수 있다.

신뢰할 수 있는 특성 사이에 상충관계가 발생하는 경우 측정은 관리 결정을 알리는 추적 가능한 기반을 제공한다. 옵션에는 재보정, 영향 완화 또는 설계, 개발, 생산 또는 사용에서 시스템 제거는 물론 다양한 보상, 탐지, 억제, 지시 및 복구 제어가 포함된다.

MEASURE 기능을 완료한 이후 측정 기준, 방법을 포함한 객관적이고 반복 또는 확장 가능한 테스트, 평가, 확인 및 검증(TEVV) 프로세스가 마련되어 문서화된다. 측정항목과 측정방법은 과학적, 법적, 윤리적 기준을 준수해야 하며 개방적이고 투명한 과정이어야 한다. 새로운 유형의 정성적, 정량적 측정을 개발해야 할 수도 있다. 각 측정 유형이 AI 리스크 평가에 고유하고 의미 있는 정보를 제공하는 정도를 고려해야 한다. 프레임워크 사용자는 시스템 신뢰성을 종합적으로 평가하고, 기존 및 긴급 리스크를 식별 및 추적하며, 측정항목의 유효성을 확인하는 능력을 향상시킨다. 측정 결과는 리스크 모니터링 및 대응 노력을 지원하기 위해 관리에 활용된다. 시간이 지나면 지식, 방법론, 리스크 및 영향이 발전하게되고 이에 따라 AI 시스템에 MEASURE 기능을 계속 적용하는 것은 프레임워크 사용자의 의무이다.

AI 리스크 측정과 관련된 실무는 NIST AI RMF 플레이북에 설명되어 있다. 표 3에는 MEASURE 기능의 카테고리와 하위 카테고리가 나열되어 있다.

표 3 : MEASURE 기능을 위한 카테고리 및 하위 카테고리

카테고리	하위 카테고리
MEASURE 1: 적절한 방법 및 지표 식별 및 적용	MEASURE 1.1: MAP 기능 중 AI 리스크 측정을 위한 접근 방식 및 측정 기준은 가장 중요한 AI 리스크부터 시작하여 선택된다. 측정되지 않거나 할 수 없는 리스크 또는 신뢰성 특성은 적절하게 문서화된다. MEASURE 1.2: 영향을 받는 지역 사회에 대한 오류 및 잠재적 영향에 대한 보고를 포함하여 AI 지표의 적절성과 기존 제어의 효율성을 정기적으로 평가하고 업데이트한다.

표 3 : MEASURE 기능을 위한 카테고리 및 하위 카테고리(계속)

카테고리	하위 카테고리
	MEASURE 1.3: 시스템의 일선 개발자 및 독립 평가자로 활동하지 않은 내부 전문가가 정기적 평가 및 업데이트에 참여한다. AI 시스템을 개발하거나 배포한 팀 외부의 도메인 전문가, 사용자, AI 행위자 및 영향을 받는 커뮤니티는 조직의 리스크 허용 범위에 따라 필요한 평가를 지원하기 위해 컨설팅을 받는다.
MEASURE 2: AI 시스템은 신뢰할 수 있는 특성으로 평가	MEASURE 2.1: TEVV 중에 사용되는 도구에 대한 테스트 세트, 측정 항목 및 세부 정보가 문서화된다. MEASURE 2.2: 인간 피험자를 대상으로 한 평가는 해당 요구사항(인간 피험자 보호 포함)을 충족하고 관련 모집단을 대표한다. MEASURE 2.3: AI 시스템 성능 또는 보증 기준은 정성적 또는 정량적으로 측정되고 배포 설정과 유사한 조건에 대해 입증된다. 조치가 문서화된다. MEASURE 2.4: MAP 기능에서 식별된 AI 시스템과 해당 구성 요소의 기능과 동작은 생산 중에 모니터링된다. MEASURE 2.5: 배포할 AI 시스템이 타당하고 신뢰할 수 있음이 입증되었다. 기술이 개발된 조건을 넘어서는 일반화 가능성의 한계가 문서화된다. MEASURE 2.6: MAP 기능에서 식별된 대로 AI 시스템의 안전 리스크를 정기적으로 평가한다. 배포될 AI 시스템은 안전한 것으로 입증되었으며, 부정적인 잔여 리스크는 리스크 허용 범위를 초과하지 않으며, 특히 지식 한계를 넘어 작동하도록 만든 경우 안전하게 실패할 수 있음이 입증되었다. 안전 지표는 시스템 신뢰성과 견고성, 실시간 모니터링, AI 시스템 오류에 대한 응답 시간을 반영한다. MEASURE 2.7: MAP 기능에서 식별된 AI 시스템 보안 및 탄력성을 평가하고 문서화한다. MEASURE 2.8: MAP 기능에서 식별된 투명성 및 책임과 관련된 리스크를 검사하고 문서화한다. MEASURE 2.9: AI 모델을 설명하고, 검증하고, 문서화하고, AI 시스템 출력을 MAP 기능에서 식별된 대로 해당 맥락 내에서 해석하여 책임 있는 사용과 거버넌스를 알린다.

표 3 : MEASURE 기능을 위한 카테고리 및 하위 카테고리(계속)

카테고리	하위 카테고리
	MEASURE 2.10: MAP 기능에서 식별된 AI 시스템의 개인 정보 보호 리스크를 검사하고 문서화한다. MEASURE 2.11: MAP 기능에서 식별된 공정성과 편견을 평가하고 결과를 문서화한다. MEASURE 2.12: MAP 기능에서 식별된 AI 모델 교육 및 관리 활동의 환경 영향과 지속 가능성을 평가하고 문서화한다. MEASURE 2.13: 측정 기능에 사용된 TEVV 측정 기준 및 프로세스의 효율성을 평가하고 문서화한다.
MEASURE 3: 시간이 지남에 따라 식별된 AI 리스크 추적 메커니즘 마련	MEASURE 3.1: 배포된 상황에서 의도된 성능과 실제 성능과 같은 요소를 기반으로 기존, 예상치 못한, 긴급 AI 리스크를 정기적으로 식별하고 추적하기 위한 접근 방식, 인력 및 문서가 마련되어 있다. MEASURE 3.2: 현재 사용 가능한 측정 기술을 사용하여 AI 리스크를 평가하기 어렵거나 측정항목을 아직 사용할 수 없는 환경에서는 리스크 추적 접근 방식을 고려한다. MEASURE 3.3: 문제를 보고하고 시스템 결과를 알리기 위한 최종 사용자와 영향을 받는 커뮤니티를 위한 피드백 프로세스가 확립되어 AI 시스템 평가 지표에 통합된다.
MEASURE 4: 측정 효율성에 대한 피드백 수집 및 평가	MEASURE 4.1: AI 리스크를 식별하기 위한 측정 접근 방식은 배포 맥락와 연결되고 도메인 전문가 및 기타 최종 사용자와의 협의를 통해 정보를 얻는다. 접근방식은 문서화된다. MEASURE 4.2: 배포 상황과 AI 생애주기 전반에 걸쳐 AI 시스템 신뢰 가치에 관한 측정 결과는 도메인 전문가 및 관련 AI 행위자의 의견을 바탕으로 시스템이 의도한 대로 일관되게 작동하는지 검증한다. 그 결과는 문서화된다. MEASURE 4.3: 영향을 받은 커뮤니티를 포함한 관련 AI 행위자와의 협의와 상황 관련 리스크 및 신뢰성 특성에 대한 현장 데이터를 기반으로 성능 개선이 측정 가능해지고 그 결과가 문서화된다.

5.4 MANAGE

MANAGE 기능은 GOVERN 기능에 의해 정의된 대로 정기적으로 보여주고 측정된 리스크에 리스크 자원을 할당한다. 리스크 처리는 사건이나 이벤트에 대한 대응, 복구 및 의사소통 계획으로 구성된다.

전문가 상담과 관련 AI 행위자(GOVERN에서 수립하고 MAP에서 수행한)의 의견을 통해 수집한 상황 정보를 이 기능에 활용하여 시스템 오류 가능성과 부정적인 영향을 줄인다. GOVERN에서 확립되고 MAP 및 MEASURE에서 활용되는 체계적인 문서화 실무는 AI 리스크관리 노력을 강화하고 투명성과 책임성을 높인다. 지속적인 개선을 위한 메커니즘과 함께 긴급 리스크를 평가하는 프로세스가 마련되어 있다.

MANAGE 기능을 완료한 후에 리스크 우선순위를 정하고 정기적인 모니터링과 개선을 위한 계획이 수립된다. 프레임워크 사용자는 배포된 AI 시스템의 리스크를 관리하고 평가되고 우선순위가 지정된 리스크를 기반으로 리스크관리 자원을 할당할 수 있는 향상된 역량을 갖추게 된다. 시간이 지남에 따라 관련 AI 행위자의 방법, 맥락, 리스크, 요구 또는 기대가 발전함에 따라 배포된 AI 시스템에 MANAGE 기능을 계속 적용하는 것은 프레임워크 사용자의 의무이다.

AI 리스크관리와 관련된 실무는 NIST AI RMF 플레이북에 설명되어 있다. 표 4에는 MANAGE 기능의 카테고리와 하위 카테고리가 나열되어 있다.

표 4 : MANAGE 기능을 위한 카테고리 및 하위 카테고리

카테고리	하위 카테고리
MANAGE 1: MAP 및 MEASURE 기능의 평가 및 기타 분석 결과를 기반으로 하는 AI 리스크의 우선순위를 정하고 이에 대응하고 관리	MANAGE 1.1: AI 시스템이 의도한 목적과 명시된 목적을 달성하는지 여부, 개발 또는 배포를 진행해야 하는지 여부에 대한 결정을 내린다. MANAGE 1.2: 문서화된 AI 리스크 처리는 영향, 가능성, 사용 가능한 자원 또는 방법을 기준으로 우선순위가 지정된다. MANAGE 1.3: MAP 기능에 의해 식별된 우선순위가 높은 것으로 간주되는 AI 리스크에 대한 대응이 개발, 계획 및 문서화된다. 리스크 대응 옵션에는 완화, 전가, 회피 또는 수용이 포함된다. MANAGE 1.4: AI 시스템의 다운스트림 인수자와 최종 사용자 모두에 대한 부정적인 잔존 리스크(완화되지 않은 모든 리스크의 합계로 정의됨)가 문서화된다.

표 4 : MANAGE 기능을 위한 카테고리 및 하위 카테고리(계속)

카테고리	하위 카테고리
MANAGE 2: AI 혜택을 극대화하고 부정적인 영향을 최소화하기 위해 관련 AI 행위자의 의견을 바탕으로 계획, 준비, 구현, 문서화 및 정보 제공	MANAGE 2.1: 잠재적 영향의 규모나 가능성을 줄이기 위해 실행 가능한 비AI 대체 시스템, 접근 방식 또는 방법과 함께 AI 리스크를 관리하는 데 필요한 자원을 고려한다. MANAGE 2.2: 배포된 AI 시스템의 가치를 유지하기 위한 메커니즘이 마련되고 적용된다. MANAGE 2.3: 이전에 알려지지 않은 리스크가 보이면 이에 대응하고 복구하기 위한 절차를 따른다. MANAGE 2.4: 의도된 사용과 일치하지 않는 성능이나 결과를 보여주는 AI 시스템을 대체, 해제 또는 비활성화하기 위한 메커니즘이 마련되고 적용되며 책임이 할당되고 이해된다.
MANAGE 3: 타사의 AI 리스크와 이점 관리	MANAGE 3.1: 타사 자원의 AI 리스크와 이점을 정기적으로 모니터링하고 리스크 제어를 적용하고 문서화한다. MANAGE 3.2: 개발에 사용되는 사전 훈련된 모델은 AI 시스템 정기 모니터링 및 유지 관리의 일부로 모니터링된다.
MANAGE 4: 식별되고 측정된 AI 리스크에 대한 대응 및 복구, 의사소통 계획을 포함한 리스크 처리 문서화 및 정기적 모니터링	MANAGE 4.1: 사용자 및 기타 관련 AI 행위자의 입력을 캡처 및 평가하는 메커니즘, 호소 및 중단, 폐기, 사고 대응, 복구 및 변경 관리를 포함하는 배포 후 AI 시스템 모니터링 계획이 구현된다. MANAGE 4.2: 지속적인 개선을 위한 측정 가능한 활동이 AI 시스템 업데이트에 통합되고 관련 AI 행위자를 포함한 이해 당사자와의 정기적인 참여가 포함된다. MANAGE 4.3: 사건과 오류는 영향을 받은 커뮤니티를 포함하여 관련 AI 행위자에게 전달된다. 사고와 오류를 추적하고, 대응하고, 복구하는 프로세스를 따르고 문서화한다.

6. AI RMF 프로파일

AI RMF 사용 사례 프로파일은 프레임워크 사용자의 요구사항, 리스크 허용 범위 및 자원을 기반으로 특정 설정 또는 애플리케이션에 대한 AI RMF 기능, 카테고리 및 하위 카테고리를 구현한 것이다. 예를 들어 AI RMF 하이어링 프로파일(hiring profile) 또는 AI RMF 페어 하우징 프로파일(fair housing profile)이 있다. 프로파일은 AI 생애주기의 다양한 단계에서 또는 특정 부문, 기술 또는 최종 사용 애플리케이션에서 리스크를 관리할 수 있는 방법을 설명하고 통찰력을 제공할 수 있다. AI RMF 프로파일은 조직이 목표에 부합하고 법적 규제 요구사항 및 모범 사례를 고려하며 리스크관리 우선순위를 반영하는 AI 리스크를 가장 잘 관리할 수 있는 방법을 결정하는 데 도움이 된다.

AI RMF 템포럴 프로파일(temporal profiles)은 특정 부문, 산업, 조직 또는 애플리케이션 맥락 내에서 특정 AI 리스크관리 활동의 현재 상태 또는 원하는 목표 상태에 대한 설명이다. AI RMF 커런트 프로파일(AI RMF Current Profile)은 현재 AI가 관리되는 방식과 현재 결과 측면에서 관련 리스크를 나타낸다. 대상 프로파일은 원하는 AI 리스크관리 목표를 달성하는 데 필요한 결과를 나타낸다.

현재 프로파일과 대상 프로파일을 비교하면 AI 리스크관리 목표를 달성하기 위해 해결해야 할 격차가 드러날 것이다. 특정 카테고리 및 하위 카테고리의 결과를 달성하기 위해 이러한 격차를 해결하기 위한 실행 계획을 개발할 수 있다. 격차를 완화시킬 우선순위는 사용자의 요구와 리스크관리 프로세스에 따라 결정된다. 또한 이 리스크 기반 접근 방식을 통해 프레임워크 사용자는 자신의 접근 방식을 다른 접근 방식과 비교하고 비용 효율적이고 우선순위가 지정된 방식으로 AI 리스크관리 목표를 달성하는 데 필요한 자원(예: 인력, 자금)을 측정할 수 있다.

AI RMF 부문 간 프로파일은 사용 사례 또는 부문 전반에 걸쳐 사용할 수 있는 모델 또는 애플리케이션의 리스크를 다룬다. 부문 간 프로파일은 대규모 언어 모델 사용, 클라우드 기반 서비스 또는 인수 등 부문 전반에 걸쳐 공통적인 활동이나 비즈니스 프로세스에 대한 리스크를 통제, 매핑, 측정 및 관리하는 방법도 다룰 수 있다. 이 프레임워크는 프로파일 템플릿을 규정하지 않으므로 구현시 유연성이 허용된다.

대전본사: (34178) 대전시 유성구 계룡로 64, 205호 / T. (042)826-6626 / F. 042-826-6627
서울지사: (05634) 서울시 송파구 가락로 252, 501호 / T. (070)4126-9583 / jklim54@daum.net