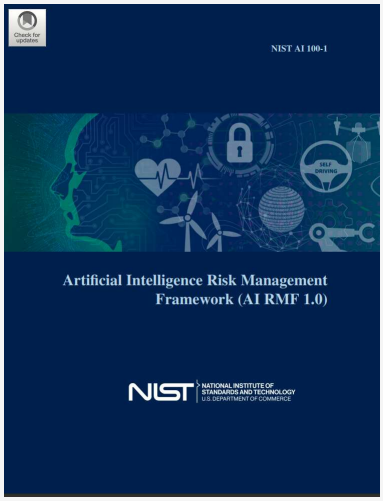
AI 리스크관리의 구조와 그 형식 (AI RMF 1.0)

K-Risk 발간편집 위원회



목차

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

※ 상기 이미지를 클릭하면 원문 다운로드가 가능합니다.

최근 챗GPT 등 생성형 AI로 인해 세상이 시끌벅적하다. AI가 가져다주는 혁신속에서 AI가 위협으로 인식되기도 한다. 유럽과 미국은 이러한 문제에 대응하기 위하여 다양한 형식의 보고서가 쏟아져 나오고 있다. 유럽에서는 AI 규제법안이 제안되기도 하였다. 국내 많은 이들도 AI에 대해 이런 저런 해석을 내놓고 있다. 2023년을 새롭게 맞이하여 K-Risk에서는 이러한 이슈에 맞는 미국 NIST에서 발간한 AI 리스크관리 프레임워크(AI RMF)를 3회에 나눠 연재 기사로 소개하고자 한다.

요약

인공지능(AI) 기술은 상업, 건강, 교통, 사이버 보안, 환경과 지구에 이르기까지 사회와 사람들의 삶을 변화시킬 수 있는 상당한 잠재력을 가지고 있다. AI 기술은 포용적 경제 성장을 주도하고 세상의 환경을 개선하는 과학적 발전을 지원할 수 있다. 그러나 AI 기술은 개인, 그룹, 조직, 커뮤니티, 사회, 환경 및 지구에 부정적인 영향을 미칠 수 있는 리스크도 내포하고 있다. 다른 유형의 기술에 대한 리스크와 마찬가지로 AI리스크는 다양한 방식으로 나타날 수 있으며 장기적 또는 단기적, 높은 확률 또는 낮은 확률, 체계적 또는 국지적, 영향력이 크거나 작은 것으로 특징지어질 수 있다.

AI RMF는 주어진 목표 집합에 대해 실제 또는 가상 환경에 영향을 미치는 예측, 권장 사항 또는 결정과 같은 출력을 생성할 수 있는 엔지니어링 또는 기계 기반 시스템으로 AI 시스템을 말한다. AI 시스템은 다양한 수준의 자율성으로 작동하도록 설계되었다.
(출처: AI에 관한 OECD 권고:2019; ISO/IEC 22989:2022).

조직이 기존 소프트웨어 또는 정보기반시스템의 리스크를 완화하는 데 도움이 되는 표준과 모범 사례는 무수히 많지만, AI 시스템으로 인해 발생하는 리스크는 여러 면에서 고유하게 나타난다(부록 B 참조). 예를 들어 AI 시스템은 시간이 지남에 따라 때로는 예기치 않게 크게 변경될 수 있는 데이터로 학습되어 이해하기 어려운 방식으로 시스템 기능과 신뢰성에 영향을 미칠 수 있다. AI 시스템과 특정상황에서의 이해관계는 종종 복잡하기 때문에 장애가 발생했을 때 이를 감지하고 대응하기가 어렵다. AI시스템은 본질적으로 사회 기술적인 특성이 있어 사회적 역학관계와 인간 행동에서 영향을 받는다. AI 리스크와 이점은 시스템 사용 방식, 다른 AI 시스템과의 상호 작용, 시스템 운영자, 시스템이 배포되는 사회적 맥락과 관련된 사회적 요인과 결합된 기술적 측면의 상호 작용에서 나타날 수 있다.

이러한 리스크로 인해 AI는 조직과 사회에서 배포하고 활용하는 데 매우 까다로운 기술이다. 적절한 통제가 없다면 AI 시스템은 개인과 커뮤니티에 불공평하거나 바람직하지 않은 결과를 증폭, 영속화 또는 악화시킬 수 있다. 그래서 적절한 통제를 통해 AI 시스템은 불공평한 결과를 완화하고 관리되어야 한다.

AI 리스크관리는 책임있는 AI시스템 개발 및 사용의 핵심 요소이다. 책임있는 AI 실무는 AI 시스템 설계, 개발 및 사용에 대한 결정을 의도된 목적과 가치에 맞게 조정하는데 도움이 될 수 있다. 책임있는 AI의 핵심개념은 인간 중심성, 사회적 책임, 지속가능성임을 강조한다. AI 리스크관리는 AI를 설계, 개발, 배포하는 조직과 내부 팀이 상황과 잠재적 영향에 대해 보다 비판적으로 생각하도록 유도하여 책임감 있는 사용과 관행을 유도할 수 있다. 또는 예상치 못한 부정적이고 긍정적인 영향에 대해 보다 비판적으로 생각하도록 유도할 수 있다. AI시스템의 리스크를 이해하고 관리하면 신뢰성을 높이고 대중의 신뢰를 구축하는 데 도움이 될 것이다.

사회적 책임이란 **"투명하고 윤리적인 행동을 통해 조직의 결정과 활동이 사회와 환경에 미치는 영향에 대한 조직의 책임 "**을 의미한다(ISO 26000:2010). 지속 가능성은 "미래 세대가 자신의 필요를 충족할 수 있는 능력을 손상시키지 않으면서 현재의 필요를 충족시키는 환경, 사회 및 경제적 측면을 포함한 글로벌 시스템 상태"를 뜻한다(ISO/IEC TR 24368:2022). 책임감 있는 AI란 공평하고 책임감 있는 기술을 의미한다. 조직의 관행이 "직업적 책임"에 따라 수행되기를 기대하면서 ISO는 "AI 시스템 및 애플리케이션 또는 AI 기반 제품이나 시스템을 설계, 개발 또는 배포하는 전문가가 사람, 사회 및 AI의 미래에 영향을 미칠 수 있는 고유한 위치를 인식하도록 하는 것을 목표로 하는 접근 방식"으로 정의한다(ISO/IEC TR 24368:2022).

2020년 국가 인공지능 개혁법안(National Artificial Intelligence Initiative Act of 2020, P.L. 116-283)에 따라 AI RMF의 목표는 AI 시스템을 설계, 개발, 배포 또는 사용하는 조직에 역량을 제공하여 AI의 여러 리스크를 관리하고 신뢰할 수 있고 책임감 있는 AI시스템 개발 및 사용을 촉진하는 것이다. 이 구성형식은 자발적이고, 권리를 보호하며 부문과 사용 사례에 구애받지 않도록 설계되어 모든 규모와 모든 부문, 사회 전반의 조직이 구성형식의 접근 방식을 구현할 수 있도록 유연성을 제공한다.

이 구성형식은 조직과 개인(본문에서는 AI 행위자라고 함)에게 이 AI시스템의 신뢰성을 높이는 접근 방식을 갖추고 시간이 지남에 따라 책임감 있는 AI 시스템의 설계, 개발, 배포 및 사용을 촉진하도록 고안되었다. 경제협력개발기구(OECD)는 AI행위자를 "AI를 배포하거나 운영하는 조직과 개인을 포함하여 AI시스템 활성주기에서 적극적인 역할을 하는 사람"으로 정의한다[OECD (2019) Artificial Intelligence inSociety—OECD iLibrary] (부록 A 참조).

AI RMF는 AI기술이 계속 발전함에 따라 AI환경에 적응하고, 다양한 수준과 역량을 가진 조직에서 운영될 수 있도록 실용적이며 사회가 AI의 혜택을 누리는 동시에 잠재적 위험으로부터 보호받을 수 있도록 하기 위한 것이다.

그 구성형식 및 지원 자원은 진화하는 기술, 전 세계의 표준 환경, AI 커뮤니티의 경험 및 반응을 기반으로 업데이트되고 확장되며 개선될 것이다. NIST는 AI RMF 및 관련 지침을 해당 국제 표준, 지침 및 관행에 계속 맞춰나갈 것이다. AI RMF가 활용됨에 따라 향후 업데이트 및 추가 자원에 대한 정보를 제공하기 위해 추가 학습이 이루어질 것이다.

그 구성형식은 두 부분으로 나뉜다. 1부에서는 조직이 AI와 관련된 리스크를 어떻게 구성할 수 있는지 논의하고 대상에 대해 설명한다. 다음으로 AI 리스크와 신뢰성을 분석하여 유효하고 신뢰할 수 있으며, 안전하고 탄력적이며, 책임있고 투명하고, 설명 가능하고 해석 가능하며, 개인정보 보호가 강화되고, 유해한 편향이 관리되어 공정해야 하는 등 신뢰할 수 있는 AI시스템의 특성을 간략하게 설명한다.

2부는 그 구성형식의 **"핵심"**으로 구성된다. 여기에서는 조직이 실제로 AI 시스템의 리스크를 해결하는 데 도움이 되는 네 가지 특정 기능을 설명한다. 이러한 기능(GOVERN, MAP, MEASURE 및 MANAGE)은 범주 및 하위 범주로 더 세분화된다. GOVERN은 조직의 AI리스크 관리 과정 및 절차의 모든 단계에 적용되지만 MAP, MEASURE 및 MANAGE 기능은 AI 시스템별 상황과 AI 수명 주기의 특정 단계에 적용될 수 있다.

구성형식과 관련된 추가적 역량요소는 NIST AI RMF 웹사이트를 통해 제공되는 AI RMF 플레이북에 포함되어 있다(https://www.nist.gov/itl/ai-risk-management-framework).

NIST가 민간 및 공공 부문과 협력하여 AI RMF를 개발하는 것은 2020년 국가 인공지능 개혁법안(National AI Initiative Act of 2020), 인공 지능에 관한 국가 안보 위원회(National Security Commission on Artificial Intelligence) 권장 사항, 그리고 기술 표준 및 관련 도구 개발에 대한 연방 참여기획(Plan for Federal Engagement in Developing Technical Standards and Related Tools)에 의한 것이었다.

이 구성형식이 개발되는 동안 공식적인 정보 요청에 대한 응답, 세 차례의 광범위한 워크숍 참석, 컨셉 페이퍼와 두 차례의 구성형식 초안에 대한 공개 의견수렴, 여러 공개 포럼에서의 토론, 여러 소그룹 회의 등을 통해 AI 커뮤니티와의 참여는 AI RMF 1.0의 개발과 NIST 및 기타 기관에서 수행한 AI 연구개발 및 평가에 영향을 미쳤다. 이 구성형식을 향상시킬 우선순위 연구 및 추가 지침은 NIST와 보다 넓은 커뮤니티가 기여할 수 있는 관련 AI 리스크관리 구성형식의 지침에 포함될 것이다.

Part 1 : 기본 정보

1. 리스크 구성과 형식

AI 리스크 관리는 시민의 자유와 권리에 대한 위협과 같은 AI 시스템의 잠재적인 부정적 영향을 최소화하는 동시에 긍정적 영향을 극대화할 기회를 제공한다. AI 리스크와 잠재적인 부정적 영향을 효과적으로 해결하고 문서화하고 관리하면 보다 신뢰할 수 있는 AI 시스템을 구축할 수 있다.

1.1 리스크에 대한 이해와 대처, 영향 및 피해

AI RMF의 맥락에서 리스크는 사고 발생 가능성과 해당 사고의 결과와 그 규모 또는 정도를 종합적으로 측정한 것을 의미한다. AI 시스템의 영향 또는 결과는 긍정적이거나 부정적이거나, 혹은 두 가지 모두의 경우일 수 있으며 기회 또는 위협이 될 수 있다(출처 : ISO 31000:2018). 잠재적 사고의 부정적인 영향을 고려할 때 리스크는 상황이나 사건이 발생할 경우 일어날 수 있는 1) 부정적 영향 또는 피해의 크기와 2) 발생 가능성의 함수이다(출처: OMB Circular A- 130:2016). 부정적 영향이나 피해는 개인, 그룹, 커뮤니티, 조직, 사회, 환경 및 지구에 미칠 수 있다.

"리스크 관리란 리스크와 관련하여 조직을 지휘하고 통제하기 위한 통합(조화)된 활동을 의미한다" (출처: ISO 31000:2018).

리스크관리 진행과 흐름은 일반적으로 부정적인 영향을 다루지만, 이 구성형식은 AI시스템의 예상되는 부정적인 영향을 최소화하고 긍정적인 영향을 극대화할 수 있는 기회를 식별하기 위한 접근방식을 제공한다. 잠재적 피해 리스크를 효과적으로 관리하면 AI 시스템을 더욱 신뢰할 수 있게 되고 사람(개인, 커뮤니티, 사회), 조직, 시스템/생태계에 잠재적 혜택을 제공할 수 있다. 리스크관리를 통해 AI개발자와 사용자는 영향을 이해하고 모델과 시스템에 내재된 한계와 불확실성을 설명할 수 있으며, 이를 통해 전반적인 시스템 성능과 신뢰성을 개선하고 AI 기술이 유익한 방식으로 사용될 가능성을 높일 수 있다.

AI RMF는 새로운 리스크가 발생하면 이에 대응할 수 있도록 설계되어있다. 이러한 유연성은 영향을 쉽게 예측할 수 없고 그 적용이 진화하는 상황에서 특히 중요하다. 일부 AI 리스크와 그 이점은 잘 알려져 있지만 부정적인 영향과 피해 정도를 평가하는 것은 어려울 수 있다. 그림 1은 AI 시스템과 관련될 수 있는 잠재적 피해의 예를 보여준다.

AI 리스크관리 노력은 인간이 모든 환경에서 AI 시스템이 작동하고 또 잘 작동된다고 가정할 수 있다는 점을 고려해야 한다. 예를 들어, 옳든 그르든 AI 시스템은 간혹 인간보다 더 객관적이거나 일반 소프트웨어보다 더 뛰어난 기능을 제공하는 것으로 인식되는 경우가 많다.

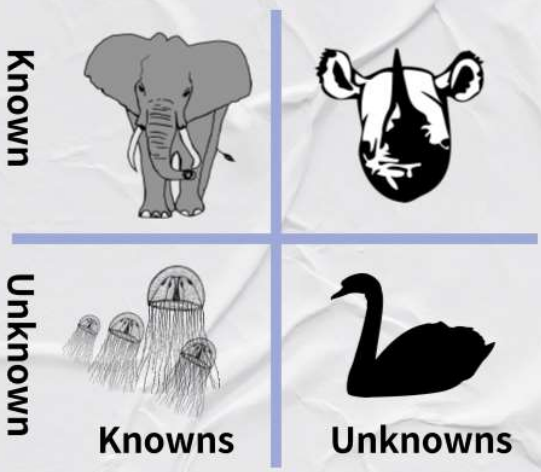
사람에 대한 피해	조직에 대한 피해	생태계에 대한 피해
• 개인 : 개인의 시민적 자유, 권리, 신체적 또는 심리적 안전 또는 경제적 기회에 대한 위해. • 집단/공동체 : 인구 하위 집단에 대한 차별 등 집단에 대한 피해. • 사회적 : 민주적 참여나 교육적 접근에 해를 끼침.	• 조직의 업무 운영에 피해를 주는 행위. • 보안 침해 또는 금전적 손실로 인한 조직의 피해. • 조직의 명예를 실추시키는 행위.	• 상호 연결되고 상호 의존적인 요소와 자원에 피해. • 글로벌 금융 시스템, 공급망 또는 상호 관련 시스템에 피해. • 천연 자원, 환경 및 지구에 피해.

그림 1. AI 시스템과 관련된 잠재적 피해의 예. 신뢰할 수 있는 AI 시스템과 책임감 있는 사용은 부정적 리스크를 완화하고 사람, 조직 및 생태계에 혜택을 줄 수 있다.

KRS
K-Risk 한국리스크전문가협의회

2022 제4회 PRM(Project Risk Management) Conference

"리스크를 경청하는 문화의 시작"



공동주최 K-Risk(한국리스크전문가협의회), 한국건설관리학회(충청지회)

일시 및 장소 2023.11.24 (목)~25(금)
(제주시 구좌읍 소재 비자숲힐링센터 대회의실)
(단체 힐링체험프로그램 포함)

참여방법 참가신청서 작성 및 메일 송부
사전참가신청 기한 2023년 10월 24일
발표참가신청 기한 2023년 10월 24일

프로그램 추후 블로그 통해 공지
대회 상세프로그램과 참가자 지원은 K-Risk의 공식 블로그 (https://k-risk.tistroy.com/81) 참고하세요

참여대상 기업/프로젝트 리스크관리 (RM) 관심자

상세한 프로그램은 우측 홈페이지나 우측 하단 QR코드를 스캔 또는 하단의 대회공식홈페이지로 방문하세요

문 의 일반참가문의 김상태간사 fasko27@pusan.ac.kr | 010-6310-6454
발표참가문의 김재영위원장 jykim@lofa.or.kr | 010-6674-9965

참가자혜택 사전참가자 및 현장(오프)참석자에게 컨퍼런스 발표집 1권 무료 제공
한국VE연구원 CRS(리스크전문가)/CVP(VE전문가) 재인증을 위한 CP점수 제공
PMI공인 PDU 8시간 제공(PDU Claim Code XXXX 제공)

후원기관 후원업체 및 기관을 모집중입니다.

참가등록자께서는 줌링크 등 상세프로그램 아래 사이트로 접근하십시오
비밀번호를 제공받지 못한 등록자께서는 사무국으로 문의하시기 바랍니다

https://k-risk.tistory.com/81

(34178) 대전시 유성구 계룡로 64, 205호 K-Risk(한국리스크전문가협의회) / T. 042-826-6626/ F. 042-826-6627