Check for updates

NIST AI 100-1



Artificial Intelligence Risk Management Framework (AI RMF 1.0)

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST AI 100-1

인공지능 위험관리 프레임워크(AI RMF 1.0)

이 간행물은 https://doi.org/10.6028/NIST.AI.100-1 에서 무료로 받아 볼 수 있다.

2023년 1월

미국 상무부 장관 Gina M. Raimondo

미국 국립표준기술연구소(National Institute of Standards and Technology) NIST 소장 겸 상무부 표준기술 담당 차관 Laurie E. Locascio

이 문서에서는 실험 절차나 개념을 적절히 설명하기 위해 특정 상업 단체, 장비 또는 자재를 언급할 수 있다. 이러한 언급은 National Institute of Standards and Technology의 추천이나 보증을 뜻하지 않으며, 해당 단체·자재·장비가 반드시 그 목적에 가장 적합하다는 뜻도 아니다.

이 간행물은 https://doi.org/10.6028/NIST.AI.100-1 에서 무료로 받아 볼 수 있다.

갱신 일정 및 버전

인공지능 위험관리 프레임워크(AI RMF)는 지속적으로 갱신되는 살아 있는 문서를 지향한다.

NIST는 갱신이 적절한지 판단하기 위해 프레임워크의 내용과 유용성을 정기적으로 검토할 것이며, AI 커뮤니티의 공식 의견을 반영한 검토는 늦어도 2028년까지 이루어질 예정이다. 프레임워크는 주요 변경과 경미한 변경을 추적·식별하기 위해 두 자리 숫자로 된 버전 체계를 사용한다. 첫 번째 숫자는 AI RMF와 그 부속 문서의 세대(예: 1.0)를 나타내며 주요 개정이 있을 때만 바뀐다. 경미한 개정은 세대 번호 뒤에 “.n”을 붙여 추적한다(예: 1.1). 모든 변경 사항은 버전 번호, 변경 일자, 변경 내용 설명을 포함한 이력을 담은 버전 관리표(Version Control Table)로 추적한다. NIST는 AI RMF Playbook을 자주 갱신할 계획이다. AI RMF Playbook에 대한 의견은 언제든지 AIframework@nist.gov 로 이메일을 보내 제출할 수 있으며, 제출된 의견은 반기마다 검토하여 반영한다.

목차

요약 1

Part 1: 기초 정보 4

1 위험의 개념 정립 **4**
- 1.1 위험, 영향 및 피해의 이해와 대응 4
- 1.2 AI 위험관리의 과제 5
  - 1.2.1 위험 측정 5
  - 1.2.2 위험 허용 수준 7
  - 1.2.3 위험 우선순위 설정 7
  - 1.2.4 위험의 조직 내 통합과 관리 8

2 대상 독자 **9**

3 AI 위험과 신뢰성 **12**
- 3.1 타당성과 신뢰성 13
- 3.2 안전성 14
- 3.3 보안성과 회복력 15
- 3.4 책임성과 투명성 15
- 3.5 설명 가능성과 해석 가능성 16
- 3.6 프라이버시 강화 17
- 3.7 공정성 – 유해한 편향의 관리 17

4 AI RMF의 효과성 **19**

Part 2: 코어와 프로파일 **20**

5 AI RMF 코어 **20**
- 5.1 Govern(거버넌스) 21
- 5.2 Map(매핑) 24
- 5.3 Measure(측정) 28
- 5.4 Manage(관리) 31

6 AI RMF 프로파일 **33**

Appendix A: 그림 2 및 3의 AI 행위자 업무 설명 **35**

Appendix B: AI 위험과 전통적 소프트웨어 위험의 차이 **38**

Appendix C: AI 위험관리와 인간-AI 상호작용 **40**

Appendix D: AI RMF의 특성 **42**

표 목록

표 1 GOVERN 기능의 범주와 하위 범주 22
표 2 MAP 기능의 범주와 하위 범주 26
표 3 MEASURE 기능의 범주와 하위 범주 29
표 4 MANAGE 기능의 범주와 하위 범주 32

그림 목록

그림 1 AI 시스템과 관련된 잠재적 피해의 예. 신뢰할 수 있는 AI 시스템과 그 책임 있는 사용은 부정적 위험을 완화하고 사람, 조직, 생태계에 이익을 가져다줄 수 있다. 5

그림 2 AI 시스템의 수명주기와 핵심 차원. OECD (2022) OECD Framework for the Classification of AI systems — OECD Digital Economy Papers를 수정하였다. 안쪽 두 원은 AI 시스템의 핵심 차원을, 바깥 원은 AI 수명주기 단계를 나타낸다. 위험관리 노력은 응용 맥락에서 계획 및 설계(Plan and Design) 기능과 함께 시작하여 AI 시스템 수명주기 전체에서 수행하는 것이 이상적이다. 대표적인 AI 행위자는 그림 3을 참조한다. 10

그림 3 AI 수명주기 단계별 AI 행위자. 시험, 평가, 검증 및 확인(testing, evaluation, verification, and validation) 작업에 관한 세부 내용을 포함하여 AI 행위자 작업에 대한 자세한 설명은 Appendix A를 참조한다. AI 모델 차원(그림 2)의 AI 행위자는 모범 관행에 따라 분리되어 있어, 모델을 구축하고 사용하는 행위자와 모델을 검증하고 확인하는 행위자가 나뉜다는 점에 유의한다. 11

그림 4 신뢰할 수 있는 AI 시스템의 특성. 타당성 및 신뢰성(Valid & Reliable)은 신뢰성의 필요조건이므로 다른 신뢰성 특성의 토대로 표시하였다. 책임성 및 투명성(Accountable & Transparent)은 다른 모든 특성과 관련되므로 세로 상자로 표시하였다. 12

그림 5 기능(function)은 AI 위험관리 활동을 최상위 수준에서 조직하여 AI 위험을 거버넌스(govern), 매핑(map), 측정(measure), 관리(manage)한다. 거버넌스는 나머지 세 기능에 정보를 제공하고 그 전반에 스며드는 교차 기능으로 설계되었다. 20

요약

인공지능(AI) 기술은 상업과 보건에서 교통과 사이버보안, 나아가 환경과 지구에 이르기까지 사회와 사람들의 삶을 변화시킬 상당한 잠재력이 있다. AI 기술은 포용적 경제 성장을 이끌고 세상의 여건을 개선하는 과학적 진보를 뒷받침할 수 있다. 그러나 AI 기술은 개인, 집단, 조직, 공동체, 사회, 환경, 지구에 부정적 영향을 줄 수 있는 위험도 초래한다. 다른 유형의 기술이 지닌 위험과 마찬가지로 AI 위험도 다양한 방식으로 나타날 수 있으며, 장기적이거나 단기적, 발생 확률이 높거나 낮음, 체계적이거나 국지적, 영향이 크거나 작음 등으로 특징지을 수 있다.

AI RMF는 AI 시스템을 주어진 목표 집합에 대해 실제 또는 가상 환경에 영향을 미치는 예측, 추천, 결정 같은 출력을 생성할 수 있는 공학적 시스템 또는 기계 기반 시스템으로 지칭한다. AI 시스템은 다양한 수준의 자율성으로 작동하도록 설계된다(출처: OECD Recommendation on AI:2019; ISO/IEC 22989:2022를 수정).

조직이 전통적인 소프트웨어나 정보 기반 시스템의 위험을 완화하도록 돕는 표준과 모범 관행은 무수히 많지만, AI 시스템이 초래하는 위험은 여러 면에서 고유하다 (Appendix B 참조). 예를 들어 AI 시스템은 시간에 따라, 때로는 상당히 그리고 예기치 않게 변할 수 있는 데이터로 학습될 수 있으며, 이는 이해하기 어려운 방식으로 시스템의 기능과 신뢰성에 영향을 미친다. AI 시스템과 그것이 배치되는 맥락은 흔히 복잡하여, 실패가 발생했을 때 이를 탐지하고 대응하기가 어렵다. AI 시스템은 본질적으로 사회기술적(socio-technical) 성격을 띠는데, 이는 사회적 역학과 인간 행동의 영향을 받는다는 뜻이다. AI의 위험과 편익은 기술적 측면이 사회적 요인, 즉 시스템이 어떻게 사용되는지, 다른 AI 시스템과 어떻게 상호작용하는지, 누가 운영하는지, 어떤 사회적 맥락에 배치되는지와 결합하여 상호작용하는 데서 나타날 수 있다.

이러한 위험 때문에 AI는 조직에서나 사회 안에서나 배치하고 활용하기가 유달리 까다로운 기술이다. 적절한 통제가 없으면 AI 시스템은 개인과 공동체에 불공평하거나 바람직하지 않은 결과를 증폭하거나 지속시키거나 악화할 수 있다. 적절한 통제가 있으면 AI 시스템은 불공평한 결과를 완화하고 관리할 수 있다.

AI 위험관리는 AI 시스템의 책임 있는 개발과 사용의 핵심 요소이다. 책임 있는 AI 관행은 AI 시스템의 설계, 개발, 사용에 관한 결정을 의도한 목적 및 가치와 일치시키는 데 도움이 될 수 있다. 책임 있는 AI의 핵심 개념은 인간 중심성, 사회적 책임, 지속가능성을 강조한다. AI 위험관리는 AI를 설계, 개발, 배치하는 조직과 그 내부 팀이 맥락과 잠재적이거나 예기치 않은 부정적·긍정적 영향을 더 비판적으로 생각하도록 촉구함으로써 책임 있는 사용과 관행을 이끌 수 있다. AI 시스템의 위험을 이해하고 관리하면 신뢰성을 높이고, 나아가 공공의 신뢰를 기르는 데 도움이 될 것이다.

사회적 책임은 조직이 “투명하고 윤리적인 행동을 통해 자신의 결정과 활동이 사회와 환경에 미치는 영향에 대해” 지는 책임을 가리킬 수 있다(ISO 26000:2010). 지속가능성은 “환경적, 사회적, 경제적 측면을 포함하여, 미래 세대가 자신의 필요를 충족할 능력을 손상하지 않으면서 현재의 필요가 충족되는 지구 시스템의 상태”를 가리킨다(ISO/IEC TR 24368:2022). 책임 있는 AI는 공평하고 책임성 있는 기술로 귀결되도록 하는 것을 의도한다. 조직의 관행은 “직업적 책임”에 따라 수행될 것으로 기대되는데, ISO는 이를 “AI 시스템과 응용 또는 AI 기반 제품이나 시스템을 설계, 개발, 배치하는 전문가가 사람, 사회, AI의 미래에 영향력을 행사할 수 있는 자신의 고유한 위치를 인식하도록 보장하는 것을 목표로 하는” 접근법으로 정의한다 (ISO/IEC TR 24368:2022).

National Artificial Intelligence Initiative Act of 2020 (P.L. 116-283)의 지시에 따라, AI RMF의 목표는 AI 시스템을 설계, 개발, 배치, 사용하는 조직에 자원을 제공하여 AI의 여러 위험을 관리하고 AI 시스템의 신뢰할 수 있고 책임 있는 개발과 사용을 촉진하도록 돕는 것이다. 본 프레임워크는 자발적이고, 권리를 보호하며, 특정 부문에 국한되지 않고, 사용 사례에 구애받지 않도록 의도되었으며, 모든 규모와 모든 부문의 조직, 그리고 사회 전반이 프레임워크의 접근법을 유연하게 실행할 수 있도록 한다.

본 프레임워크는 조직과 개인, 즉 여기서 AI 행위자(AI actor)라 부르는 이들에게 AI 시스템의 신뢰성을 높이는 접근법을 갖추게 하고, 시간이 흐르면서 AI 시스템의 책임 있는 설계, 개발, 배치, 사용을 촉진하도록 돕기 위해 설계되었다. 경제협력개발기구(OECD)는 AI 행위자를 “AI를 배치하거나 운영하는 조직과 개인을 포함하여 AI 시스템 수명주기에서 능동적 역할을 하는 이들”로 정의한다 [OECD (2019) Artificial Intelligence in Society—OECD iLibrary](Appendix A 참조).

AI RMF는 실용적이고, AI 기술이 계속 발전함에 따라 AI 지형에 적응하며, 조직이 다양한 정도와 역량으로 운용할 수 있도록 하여, 사회가 AI의 혜택을 누리면서도 그 잠재적 피해로부터 보호받을 수 있도록 하는 것을 의도한다.

본 프레임워크와 지원 자원은 진화하는 기술, 전 세계의 표준 지형, AI 커뮤니티의 경험과 피드백을 바탕으로 갱신, 확장, 개선될 것이다. NIST는 AI RMF와 관련 지침을 적용 가능한 국제 표준, 지침, 관행과 계속 일치시켜 나갈 것이다. AI RMF가 사용됨에 따라 추가로 얻는 교훈은 향후 갱신과 추가 자원에 반영될 것이다.

본 프레임워크는 두 부분으로 나뉜다. Part 1은 조직이 AI 관련 위험을 어떻게 규정할 수 있는지 논의하고 의도한 대상 독자를 기술한다. 이어서 AI 위험과 신뢰성을 분석하고 신뢰할 수 있는 AI 시스템의 특성을 개괄하는데, 그 특성에는

타당성과 신뢰성, 안전성, 보안성과 복원력, 책임성과 투명성, 설명 가능성과 해석 가능성, 프라이버시 강화, 해로운 편향이 관리된 공정성이 포함된다.

Part 2는 프레임워크의 "핵심(Core)"으로 구성된다. 여기서는 조직이 실제로 AI 시스템의 위험에 대처하도록 돕는 네 가지 구체적 기능을 기술한다. 이 기능들, 즉 GOVERN, MAP, MEASURE, MANAGE는 다시 범주와 하위 범주로 세분된다. GOVERN은 조직의 AI 위험관리 프로세스와 절차의 모든 단계에 적용되는 반면, MAP, MEASURE, MANAGE 기능은 AI 시스템별 맥락과 AI 수명주기의 특정 단계에서 적용될 수 있다.

본 프레임워크와 관련된 추가 자원은 AI RMF Playbook에 수록되어 있으며, NIST AI RMF 웹사이트 https://www.nist.gov/itl/ai-risk-management-framework 에서 이용할 수 있다.

NIST는 민간 및 공공 부문과 협력하여 AI RMF를 개발하였으며, 이는 「2020년 국가 AI 이니셔티브법(National AI Initiative Act of 2020)」, 인공지능 국가안보위원회(National Security Commission on Artificial Intelligence)의 권고, 「기술 표준 및 관련 도구 개발을 위한 연방 참여 계획(Plan for Federal Engagement in Developing Technical Standards and Related Tools)」이 요구하는 NIST의 더 넓은 AI 활동에 따라 지시된 것이고 그와 일관된다. 공식 정보요청(Request for Information)에 대한 응답, 많은 이가 참석한 워크숍 세 차례, 개념 문서와 프레임워크 초안 두 건에 대한 공개 의견, 여러 공개 포럼에서의 논의, 수많은 소규모 그룹 회의를 통해 프레임워크 개발 기간 동안 AI 공동체와 교류하였고, 이는 AI RMF 1.0의 개발은 물론 NIST와 여러 기관이 수행하는 AI 연구개발 및 평가에도 반영되었다. 본 프레임워크를 강화할 우선 연구 과제와 추가 지침은 관련 문서인 AI 위험관리 프레임워크 로드맵(AI Risk Management Framework Roadmap)에 담길 것이며, NIST와 더 넓은 공동체가 이 로드맵에 기여할 수 있다.

Part 1: 기초 정보

1. 위험의 개념 정립

AI 위험관리는 시민적 자유와 권리에 대한 위협과 같은 AI 시스템의 잠재적 부정적 영향을 최소화하는 길을 제시하는 동시에, 긍정적 영향을 극대화할 기회도 제공한다. AI 위험과 잠재적 부정적 영향을 효과적으로 다루고 문서화하며 관리하면 더 신뢰할 수 있는 AI 시스템으로 이어질 수 있다.

1.1 위험, 영향, 피해의 이해와 대응

AI RMF에서 위험이란 어떤 사건이 발생할 확률과 그 사건이 낳는 결과의 크기 또는 정도를 종합한 척도를 말한다. AI 시스템의 영향, 즉 결과는 긍정적일 수도, 부정적일 수도, 둘 다일 수도 있으며 기회나 위협을 낳을 수 있다(Adapted from: ISO 31000:2018). 잠재적 사건의 부정적 영향을 고려할 때, 위험은 1) 해당 상황이나 사건이 발생할 경우 생길 부정적 영향, 즉 피해의 크기와 2) 발생 가능성의 함수이다(Adapted from: OMB Circular A-130:2016). 부정적 영향이나 피해는 개인, 집단, 지역사회, 조직, 사회, 환경, 그리고 지구가 겪을 수 있다.

> “위험관리란 위험과 관련하여 조직을 지휘하고 통제하기 위한 조정된 활동을 말한다”(Source: ISO 31000:2018).

위험관리 프로세스는 일반적으로 부정적 영향을 다루지만, 본 프레임워크는 예상되는 AI 시스템의 부정적 영향을 최소화하고 긍정적 영향을 극대화할 기회를 파악하는 접근법을 제시한다. 잠재적 피해의 위험을 효과적으로 관리하면 더 신뢰할 수 있는 AI 시스템으로 이어지고, 사람(개인, 지역사회, 사회), 조직, 시스템/생태계에 잠재적 혜택을 가져다줄 수 있다. 위험관리는 AI 개발자와 사용자가 영향을 이해하고 자신의 모델과 시스템에 내재한 한계와 불확실성을 고려할 수 있게 하며, 이는 다시 전반적인 시스템 성능과 신뢰성을 높이고 AI 기술이 유익한 방식으로 사용될 가능성을 높일 수 있다.

AI RMF는 새로운 위험이 나타날 때 이를 다룰 수 있도록 설계되었다. 이러한 유연성은 영향을 쉽게 예견하기 어렵고 응용이 계속 발전하는 상황에서 특히 중요하다. 일부 AI 위험과 혜택은 잘 알려져 있지만, 부정적 영향과 피해의 정도를 평가하기는 어려울 수 있다. 그림 1은 AI 시스템과 관련될 수 있는 잠재적 피해의 예를 보여 준다.

AI 위험관리 노력에서는 사람들이 AI 시스템이 모든 환경에서 작동하며, 그것도 잘 작동한다고 여길 수 있다는 점을 고려해야 한다. 예를 들어, 옳든 그르든 AI 시스템은 흔히 인간보다 더 객관적이거나 일반 소프트웨어보다 더 뛰어난 능력을 제공한다고 인식된다.



그림 1. AI 시스템과 관련된 잠재적 피해의 예. 신뢰할 수 있는 AI 시스템과 그 책임 있는 사용은 부정적 위험을 완화하고 사람, 조직, 생태계에 대한 혜택에 기여할 수 있다.

1.2 AI 위험관리의 과제

아래에서 몇 가지 과제를 설명한다. AI 신뢰성을 추구하며 위험을 관리할 때 이 과제들을 고려해야 한다.

1.2.1 위험 측정

명확히 정의되지 않았거나 충분히 이해되지 않은 AI 위험이나 실패는 정량적으로든 정성적으로든 측정하기 어렵다. AI 위험을 적절히 측정할 수 없다고 해서 그 AI 시스템이 반드시 높은 위험이나 낮은 위험을 지닌다는 뜻은 아니다. 위험 측정의 과제로는 다음과 같은 것들이 있다.

제3자 소프트웨어, 하드웨어, 데이터 관련 위험: 제3자 데이터나 시스템은 연구개발을 가속하고 기술 이전을 촉진할 수 있다. 그러나 위험 측정을 복잡하게 만들 수도 있다. 위험은 제3자 데이터, 소프트웨어, 하드웨어 그 자체에서도, 그것이 사용되는 방식에서도 생길 수 있다. AI 시스템을 개발하는 조직이 쓰는 위험 지표나 방법론은 그 시스템을 배포하거나 운영하는 조직이 쓰는 위험 지표나 방법론과 맞지 않을 수 있다(역주: 원문 「uses」는 「used」의 오기로 보인다). 또한 AI 시스템을 개발하는 조직이 자신이 사용한 위험 지표나 방법론을 투명하게 공개하지 않을 수도 있다. 특히 내부 거버넌스 구조와 기술적 안전장치가 충분하지 않은 경우, 고객이 제3자 데이터나 시스템을 AI 제품이나 서비스에 사용하거나 통합하는 방식에 따라 위험 측정과 관리가 복잡해질 수 있다. 그럼에도 모든 당사자와 AI 행위자는 자신이 개발하거나 배포하거나 사용하는 AI 시스템의 위험을, 그것이 독립형 구성요소이든 통합된 구성요소이든 관리해야 한다.

새로 나타나는 위험의 추적: 새로 나타나는 위험을 식별하고 추적하며 이를 측정할 기법을 검토하면 조직의 위험관리 노력이 강화될 것이다.