The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 24, No. 1, pp.115-123, Feb. 29, 2024. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2024.24.1.115
JIIBC 2024-1-18

인공지능과 위험관리에 대한 사례 연구 - RAI Toolkit을 중심으로

Case Study on Artificial Intelligence and Risk Management - Focusing on RAI Toolkit

신선영*

Sunyoung Shin*

요 약 본 연구의 목적은 인공지능과 위험관리라는 2가지 키워드를 통해 어떻게 인공지능 서비스의 장점 활용과 한계요인을 동시에 극복하는데 기여 하고자 한다. 이를 위해 2가지 사례인 (1) 인공지능을 활용한 위험 모니터링 프로세스 제시와 (2) 인공지능 서비스의 개발 및 운영에서 등장하는 한계요인을 최소화하기 위한 운영 툴킷에 대해 소개 하였다. 이 사례 분석을 통해 다음과 같은 시사점이 제안하고자 한다. 첫째, 인공지능 서비스는 우리 삶에 깊숙이 관여하고 있으며 이로 인해 등장하는 한계 요인을 최소화하는 장치가 필요하다. 둘째, 인공지능을 활용한 위험관리 모니터링은 적합하고 신뢰성이 있는 데이터 확보가 우선적으로 고려되어야 한다. 셋째, 인공지능 서비스의 개발과 운영시 등장하는 한계를 극복하기 위해서는 업무 단계별로 위험관리 프로세스를 적용하여 상시 모니터링이 요구된다 라는 것이다. 본 연구는 발전하고 있는 인공지능이 제공하고 한계요인을 최소화 할 수 있는 방안에 대한 연구이며 향후 관련 시장의 성장과 발달에서 위험관리에 대한 연구에 기여 할 수 있다.

Abstract The purpose of this study is to contribute to how the advantages of artificial intelligence (AI) services and the associated limitations can be simultaneously overcome, using the keywords AI and risk management. To achieve this, two cases were introduced: (1) presenting a risk monitoring process utilizing AI and (2) introducing an operational toolkit to minimize the emerging limitations in the development and operation of AI services. Through case analysis, the following implications are proposed. First, as AI services deeply influence our lives, the process are needed to minimize the emerging limitations. Second, for effective risk management monitoring using AI, priority should be given to obtaining suitable and reliable data. Third, to overcome the limitations arising in the development and operation of AI services, the application of a risk management process at each stage of the workflow, requiring continuous monitoring, is essential. This study is a research effort on approaches to minimize limitations provided by advancing artificial intelligence (AI). It can contribute to research on risk management in the future growth and development of the related market, examining ways to mitigate limitations posed by evolving AI technologies.

Key Words : Artificial intelligence service, Risk management, AI limitted factors, Evaluation tool, RAI tool kit

*정회원, 한국지능정보사회진흥원 AI데이터 활용팀
접수일자 2023년 12월 30일, 수정완료 2024년 1월 28일
게재확정일자 2024년 2월 9일

Received: 30 December, 2023 / Revised: 28 January, 2024 / Accepted: 9 February, 2024
*Corresponding Author: shinsy@nia.or.kr
Department of AI Data, National Information Society Agency, Korea

I. 서 론

AI 기술은 딥러닝, 신경 학습, 네트워크, 바이너리, 양자 컴퓨팅 등 다양한 형태로 나타나고 있다. 이러한 기술을 통해 AI는 도르래, 증기 기관, 컴퓨터와 같이 인간의 능력을 강화하고 사회 복지를 향상시키는 도구로 역할을 하고 있다.

업무 효율성을 극대화 시켜주는 인공지능 서비스도 다양한 한계 요인을 제공하고 있으며 이를 워험 관리 측면에서 검토하고 있다. 시장에 제공되는 다양한 인공지능 서비스의 장점을 활용하면서 단점을 줄이기 위한 방안에 대해 논의가 이루어지고 있다.

인공지능을 정책 수립 도구로서 사용되고 있으며 이로 인해 인공지능 서비스의 한계 요인 및 위험이 상존하고 있다. 인공지능이 제공하는 업무의 효율성으로 민간 서비스 뿐만 아니라 공공 서비스에 적용되고 있다. 특히, 공공정책 수립시에 도입이 되고 있으며 이로 인해 상존하는 위험을 제거하려는 다양한 시도가 일어나고 있다.

본 연구는 인공지능과 위험관리라는 키워드를 중심으로 2가지 사례를 분석하고 있다. 첫 번째 사례는 인공지능을 활용하여 시장에 등장하는 위험 요인을 추적하는 방법에 대한 사례를 분석하고 있다. 그리고 또 다른 사례는 제공되는 인공지능 서비스의 한계 요인을 추적하고 개발과 운영 단계에서 이를 제거하여 위험 요인을 최소화하는 방안에 대한 사례이다.

본 연구 결과는 인공지능을 활용한 정책 수립시 등장하는 한계요인을 제거하고 인공지능 기반 정책 수립 관련 등장하는 위험 요소를 제거하는 사례로서 정책적 시사점을 제공하는 것에 의의를 두고 있다. 따라서 관련 분야에 대한 다양한 연구가 지속적으로 이루어져야 한다.

II. 인공지능 위험관리

1. 개요

인공지능은 기계가 경험을 통해 학습하고 새로운 입력 내용에 따라 기존 지식을 조정하며 사람과 같은 방식으로 과제를 수행할 수 있도록 지원하는 기술이다[1]

인공지능은 학습, 문제 해결, 패턴 인식 등과 같이 주로 인간 지능과 연결된 인지 문제를 해결하는 데 주력하는 컴퓨터 공학 분야에 속한다[2][3].

인공지능 기술이 빠르게 발달하고 다양한 산업에서 적용되고 있으며 민간 서비스 뿐만 아니라 공공 서비스에도 접목이 이루어지고 있다. 그리고 인공지능(AI) 애플리케이션은 인류의 발전과 지속 가능한 개발 목표 달성을 위한 기회를 확보하기 위해 지속적으로 확대하고 있다

위험관리는 조직이나 프로젝트에서 발생할 수 있는 다양한 위험을 식별하고 분석하여 그 위험에 대비하기 위한 전략을 수립하고 이행하는 과정을 말한다. 위험은 불확실성으로부터 발생하며 조직이나 프로젝트의 목표를 달성하는 과정에서 예상치 못한 문제나 손실을 초래할 수 있다. 개발 또는 운영되고 있는 인공지능 서비스에도 위험관리는 중요한 개념이다.

2. 위험관리 측면

인공지능 관련 위험관리 측면은 다음과 같은 2가지 측면을 볼 수 있다.

그림 1. 인공지능과 위험관리
Fig. 1. Artificial Intelligence and Risk Management

첫째, 우리 환경에서 등장하고 있는 위험 요소를 신속하고 효율적으로 추정하여 보고하는 영역이다. 이 과정을 다양한 영역에서 활용되고 있으며 본 연구에서는 이를 위해 적용되는 일반적인 프로세스에 대해 제공하고 있다.

두 번째 영역은 인공지능을 기반으로 개발 또는 운영하고 있는 서비스에서 등장할 수 있는 위험 요소를 제거하거나 최소화하는 것이다.

여러 이유로 인해 인공지능 시스템은 예를 들면, 부정확성, 편향성, 개인 정보 보호 문제 등 다양한 위험을 내포하고 있을 수 있다. 이 영역은 여러 국가와 기관에서 민감하게 고려를 하고 있으며 일부 조직은 이미 관련 부분에 대해 개발하여 적용하고 있다. 이를 구현하기 위해 일반적으로 다음과 같은 사항들이 검토되고 있다.

표 1. 인공지능 서비스 위험관리 방향
Table 1. Artificial intelligence service risk management direction

구분	내용
투명성과 설명 가능성 강화	인공지능 시스템이 내린 결정이나 예측을 이해할 수 있어야 함 투명성과 설명 가능성을 강화함으로써 사용자들은 시스템이 어떻게 작동하는지 이해하고, 결정에 대한 신뢰를 가질 수 있음
편향성 탐지 및 개선	알고리즘의 훈련 데이터나 구현 과정에서 발생할 수 있는 편향성을 탐지하고 개선하는 것이 중요 특히 인종, 성별, 연령 등에 대한 편향을 최소화하여 공정하고 중립적인 서비스를 제공해야 함
보안 및 개인정보 보호 강화	인공지능 시스템은 민감한 정보를 취급 보안 및 개인 정보 보호에 특히 신경을 써야 함 데이터의 안전한 저장과 전송, 암호화 등의 기술적 조치를 통해 개인 정보를 보호해야 함
사용자 교육과 인지	사용자들에게 어떤 정보가 수집되고 어떻게 활용되는지에 대한 명확한 교육이 필요합니다. 이를 통해 사용자들은 서비스의 기능과 위험에 대해 더 잘 이해하고 적절한 선택을 할 수 있음
윤리적 가이드라인 준수	인공지능 개발 및 운영에는 윤리적 가이드라인을 준수 국제적이고 산업 표준을 따르면서, 사회적 책임을 다하는 개발 및 운영이 필요
피드백 및 개선 주기 구축	사용자 피드백을 수집하고, 모델 및 시스템을 지속적으로 개선하는 주기를 구축해야 함 이는 실제 운영 환경에서 발생하는 문제에 대응하고 서비스를 지속적으로 향상시키는 데 도움이 됨
규제 및 규범 준수	국가 및 산업 규제에 따라 인공지능 시스템을 개발하고 운영해야 함 규제 준수는 사용자의 안전성과 개인 정보 보호를 보장하는데 중요한 역할을 함

III 위험관리 적용 사례

1. 인공지능을 활용한 위험 모니터링

가. 위험 모니터링

인공지능 기반 위험 모니터링(AIM: Artificial Intelligence Monitoring)은 매우 중요한 영역으로 인공지능이 인간을 대신하여 업무를 종합적이고 효율적으로 분석하여 결과를 제공하는 것이다.

이와 관련하여 OECD에서는 최근 AIM을 발표하였다. AIM(Artificial Intelligence Monitoring)은 AI 사고와 AI 위험 및 재난과 같은 관련 용어를 정의하는 전문가 그룹의 작업을 통해 정보를 얻고 있다. 이와 동시에 AIM은 AI 사고의 정의와 보고 프레임워크가 실제 AI 사고와 함께 작동하는지 확인하기 위해 현실적 점검 내용을 제공하고 있다.

기본적으로 등장하는 위험을 파악하기 위해 능동 학습 기술이 적용된 딥러닝 모델을 사용한다. 또한 AIM을 위한 데이터 수집 및 분석은 AI 사고에 대한 정보의 신뢰성, 객관성 및 품질을 최대한 보장하기 위해 수행된다.

나. 모니터링 기본 조건

모니터링 기본 조건은 모니터링을 위한 정보의 투명성 공개이다. AIM 과정을 통해 우리에게 제공되는 정보의 투명한 공개를 위해 다음 사항이 검토되어야 한다.

표 2. 정보 투명성 공개를 위해 검토 사항 요소
Table 2. Considers for disclosure of information transparency

구분	내용
목표	OECD AI 사고 모니터("AIM") 사용에는 www.oecd.org/termsandconditions 에 있는 이용 약관이 적용이 됨 공개는 AIM에 포함된 정보에 대한 더 큰 투명성을 제공하는 것을 목표로 함
제3자 정보	AIM은 AI 관련 과제의 환경을 이해하기 위한 접근 가능한 출발점 역할을 함 AIM에는 OECD와 관련이 없는 다양한 제3자 매체 및 뉴스 수집 기관의 뉴스 기사가 포함되어 있다는 점을 유의해야 함
표현된 견해	AIM에 표현된 모든 견해나 의견은 이를 작성한 제3자 매체의 것임 또한 뉴스 기사나 사건이 포함되었다고 해서 OECD가 이를 지지하거나 권장하는 것은 아님
오류 및 누락	OECD는 AIM에 제공된 제3자 정보의 정확성, 완전성 또는 유효성을 보장할 수 없으며 독립적으로 검증하지 않음 AIM에 포함된 정보에는 다양한 오류와 누락이 포함될 수 있다는 점을 유의해야 함
지적 재산권	AIM에 언급되거나 인용되거나 포함된 모든 저작권, 상표, 서비스 마크, 집단 마크, 디자인 권한 또는 기타 지적 재산권이나 소유권은 해당 소유자의 재산임.

다. 모니터링 프로세스

AIM를 위해 사용하는 프로세스는 다음과 같다.

1단계는 딥러닝 모델의 조정이다. 위험을 파악하기 위해 전 세계적으로 평판이 좋은 국제 미디어에 보도된 AI 사건은 능동 학습 기술을 사용하여 미세 조정된 딥 러닝 모델을 사용하여 식별되고 분류된다.

2단계에서는 유사한 모델을 사용하여 사건의 심각도, 산업, 관련 AI 원칙, 피해 유형 및 영향을 받는 이해관계자를 포함하여 AI 시스템 분류를 위한 OECD 프레임워크의 다양한 범주로 사고를 분류한다[5].

3단계는 평판이 우수한 데이터를 확보한다. 분석을 위해 각 뉴스 기사의 제목, 초록 및 처음 몇 단락을 기반으로 수행한다. 뉴스 기사는 세계 뉴스를 모니터링 하고 매일 150,000개 이상의 영어 기사가 처리되는 뉴스 기사에 보고된 특정 이벤트 유형을 감지할 수 있는 뉴스 인텔리전스 플랫폼인 Event Registry에서 제공한다.

4단계는 분석된 내용을 기반으로 우리가 활용이 가능

한 시사점을 제공한다. 이러한 사고가 전 세계 모든 AI 사고의 일부일 뿐이라는 가능성을 인식하면서도 공개적으로 보고된 이러한 사고는 증거 기반 구축을 위한 유용한 시작점을 제공한다. 사건은 동일한 사건을 다루는 하나 이상의 뉴스 기사로 구성될 수 있다. 또한, 사건은 사건에 대해 보고된 기사 수와 의미론적 유사성에 따라 결정된 특정 검색어와의 관련성을 기준으로 정렬된다.

5단계는 추출된 정보에 대해 공개하고 이를 확인 할수 있는 링크를 생성하여 제공한다. 즉 마지막으로, 완전성을 위해 특정 사건을 보고하는 모든 기사에 대한 링크가 제공된다.

그림 2. AIM 기반 자료 제공 프로세스
Fig. 2. AIM-based data provision process

뉴스 기사의 AI 사건 정보를 보완하기 위해 공개 제출 프로세스가 활성화될 수도 있다. 보고의 일관성을 보장하기 위해 기존 분류 알고리즘을 활용하여 텍스트 제출을 처리하고 특정 사고 보고서에 대한 사전 선택 태그를 제공하고자 한다. 또한, 뉴스 기사의 사건 정보는 법원 판결 및 공공 감독 기관의 결정에 의해 보완될 수 있다.

라. 모니터링을 위한 기술

첫째, 사고 감지 이다. AI 사고를 지도 학습 작업, 특히 텍스트 분류 문제로 탐지하는 문제를 프레임화할 수 있다. 주어진 문장에서 주어진 엔터티 쌍과 관련된 AI 사고를 표현하는가를 검토한다.

문장에서 개체 쌍을 분류하려면 먼저 텍스트에 존재하는 개체를 식별해야 한다. 텍스트에서 엔터티를 식별하기 위해 spacy[6] 를 사용하는 명명된 엔터티 인식(NER)과 별도의 엔터티 감지 및 연결 시스템인 Wikifier[7] 를 모두 사용한다. 그리고 다음 AI 사고를 탐지하기 위해 기계 학습 모델을 훈련하는 데 사용할 지도 데이터 세트를 생성한다.

둘째, 모델 수립이다. 문장에서 개체 쌍을 분류하는데 사용되는 모델은 변환기[8] 신경망을 기반으로 한다. 이는 BERT와 유사한[9] 사전 학습된 언어 모델인 RoBERTa [10]를 사용하여 문장의 텍스트를 인코딩한다. 인코딩하기 전에 특수 토큰으로 분류되는 엔터티 쌍을 둘러싸도록 문장의 텍스트를 수정하고 엔터티 앞에 엔터티 유형(예: 조직, 위치, 제품 등)을 추가한다.

이는 관계 분류에 대한 자료[11]에 설명된 절차를 따른다. 모델 아키텍처 자체를 따른다[12]. 변환기는 입력 텍스트를 인코딩한다. 각 엔터티 이전의 특수 토큰에 해당하는 출력 임베딩은 단일 벡터로 함께 연결된 다음 선형 레이어로 구성된 분류 헤드로 전달된다. 그 다음에는 가능한 이벤트 클래스에 대해 정규화된 확률 분포를 생성하기 위한 Softmax 활성화 함수가 이어진다. 이는 해당 문장의 개체 쌍이 AI 사고에 해당할 확률에 해당한다.

셋째, 데이터 셋트의 확정이다. AI 사건을 탐지하는 문제는 지도형 기계 학습 작업으로 공식화된다. 학습 프로세스는 일반적으로 모델 학습으로 알려져 있다. 이는 모델이 훈련 데이터(예: "훈련 세트")의 예를 입력으로 사용하고 그로부터 학습한다는 것을 의미한다. 그런 다음 훈련된 모델의 성능은 훈련 세트에 포함되지 않은 데이터(예: "검증 세트")를 사용하여 추정된다. 검증 세트에는 긍정적인 예와 부정적인 예가 모두 포함되어 있다. 여기서 긍정적인 예는 실제 AI 사고의 예이고 부정적인 예는 AI 사고가 아닌 것으로 알려져 있다.

넷째, 성과 지표 수립이다. 어떤 성과를 기대하고 있는지에 대한 명확한 지표가 제공되어야 한다.

다섯째, 실험결과 분석이다. 상기의 준비과정을 통해 획득한 결과에 대한 분석이다.

여섯째, 사고의 분류이다. 상기의 분석내용을 기반으로 사고를 어떻게 분류할 수 있을지 제시해야 한다. 여러 차원에 따라 AI 사고를 추가로 분류한다. 여기에는 사고의 심각도 수준, 피해 유형, 영향을 받는 이해관계자를 포함한다.

심각도 수준과 피해 유형은 다중 클래스 분류 문제로 정의된다. 즉, 각 문제에 대해 값 중 하나만 참일 수 있음을 의미한다. 동일한 사고로 여러 그룹이 피해를 입을 수 있으므로 영향을 받는 이해관계자 카테고리는 다중 라벨 분류 문제로 정의되었다. 이 경우 '알 수 없음'이라는 라벨은 피해를 입은 것으로 확인된 다른 그룹이 없는 경우에만 사용된다. 잠재적, 가상, 미래 피해 또는 실현되지 않은 위험과 관련된 사건을 포괄하기 위해 추가 이진 범주인 "미래 위협"이 생성되었다.

일곱째, 확정된 모델 확정이다. 이 과정을 통해 수립된 정보와 결과를 기반으로 인공지능 기반 사고 모델에 대한 사례로 확정하여 향후에 비슷한 모델의 경우 이 방법으로 분석할 수 있는 체계가 수립된다.

2. 인공지능 서비스 위험 관리 툴킷 RAI

가. 인공지능 서비스 위험관리 도입 이유

궁극적으로 AI 시스템은 신뢰를 기반으로 할 때 적용할 수 있다. 현재 각 영역에서 수행하는 모든 일의 기반이 되는 AI에 대한 원칙적인 접근 방식을 가지고 있어야 하며 이것을 책임감 있는 AI라고 부르는데, 이를 위해 처리하는 방법에 대한 고민이 필요하다. 그리고 책임 있는 AI는 최첨단 기술이 시대를 초월한 가치를 제공할 수 있다.

미국 국방부(DOD)의 책임 있는 인공지능 전략 및 구현 경로에 명시된 바와 같이, 국방부는 AI를 설계, 개발, 테스트, 조달, 배치 및 사용할 때 합법적이고 윤리적인 행동에 대한 국방부의 확고한 의지가 적용된다는 것을 입증해야 한다.

국방부(DoD)내 RAI 구현의 일환으로 이 툴킷/평가는 사용자에게 DoD AI 기능을 책임감 있게 개발, 배포 및 사용하는 데 도움이 되는 다양한 고려 사항을 제공하는 것을 목표로 한다. 이 툴킷 기반 평가는 AI 지원 시스템의 전체 수명주기(설계, 개발, 배포 및 사용 포함)가 DoD의 AI 윤리 원칙과 일치하는지 확인하기 위한 내용을 제공한다.

이 평가 도구는 AI 수명주기의 복잡한 단계를 거치는 개인을 위한 리소스 역할을 하여 프로젝트 내에서 RAI의 중추적인 차원을 탐색할 수 있도록 지원한다. 이는 유사한 노력에서 얻은 모범적인 결과물 및 통찰력과 함께 해당 부서가 지원하는 다양한 도구를 제공함으로써 업무 효율성을 지원한다.

나. 사용자

RAI 툴킷은 사용자 커뮤니티가 운영 요구 사항을 식별하고 해당 요구 사항을 충족하는 AI 기능에 대한 운영 요구사항을 설정하는 수명주기 프로세스의 다양한 측면을 지원하도록 구축되었다.

제도적 요구 사항 소유자는 이를 기능적 요구 사항으로 변환하고 획득 커뮤니티와 협력하여 해당 요구사항을 충족하는 기능을 구축하거나 구매한다. 획득 커뮤니티는 요구사항을 기능에 대한 성능 사양으로 더 변환한다. 프로그램 관리자는 데이터 엔지니어, 모델 개발자, 사용자 경험 디자이너 등을 포함한 개발팀이 실행하는 기능 조달을 감독한다. 고위 리더는 수명주기 전반에 걸쳐 프로세스를 모니터링하고 감독한다.

AI 윤리 및 위험 전문가, 테스트 및 평가 전문가, 개인정보 보호 및 사이버 보안 담당자를 포함하여 추가 전문가도 여러 단계에 걸쳐 지원한다. 이러한 페르소나는 RAI 평가에 사용되어 해당 페르소나에 적용되는 RAI 수명주기의 관련 측면을 식별한다.

PoR(Program of Record) 경로 외에도 이 툴킷의 사용자는 식별된 운영 요구에 맞게 내부적으로 AI 기능을 개발하려는 소규모 팀(또는 한 팀으로 구성된 팀) 출신일 수 있다. 이 경우 해당 팀의 개인에게 여러 페르소나가 적용 가능하다.

사업의 규모에 관계없이 이 포괄적인 툴킷과 평가 프레임워크는 RAI를 효과적으로 운영하기 위해 고려해야 하는 중요한 차원을 확인한다. 소규모 팀이라도 이 리소스에 설명된 모든 단계와 역할을 철저히 조사하여 이러한 중추적인 측면을 적절히 고려해야 한다.

툴킷 인터페이스는 AI 프로그램의 특정 상황에 따라 맞춤화/모듈화되도록 설계되었다. RAI 평가 단계는 RAI 전략 및 구현 경로에 설명된 대로 시작부터 사용까지 전체 AI 개발 수명주기를 따른다. 이 툴킷의 사용자는 프로젝트 라이프사이클의 어느 시점에서든 자료를 참조하여 해당 단계에서 고려해야 할 관련 RAI 질문을 식별하고 해당 질문에 대한 답변을 지원하는 데 사용 가능한 도구에 액세스할 수 있다.

다. RAI 툴킷

RAI(The Responsible Artificial Intelligence: 책임 있는 인공지능) 툴킷은 RAI 모범 사례 및 DoD (Department of Defense) AI 윤리 원칙에 대한 AI 프로젝트의 조정을 식별, 추적 및 개선하는 동시에 혁신 기회를 활용하는 중앙 집중식 프로세스를 제공한다.

인공지능 기반 서비스에서 발생할 수 있는 위험관리 목적을 위해 RAI(책임 있는 인공지능) 툴킷은 AI 프로젝트를 RAI 모범 사례 및 DoD(Department of Defense) AI 윤리 원칙에 맞게 식별, 추적, 개선하는 동시에 혁신 기회를 활용하는 중앙 집중식 프로세스를 제공하고 있다.

RAI 툴킷에서 위험관리 주요 내용은 AI 제품 라이프사이클 전반에 걸쳐 맞춤형 및 모듈식 평가, 도구를 통해 사용자를 안내하는 직관적인 흐름을 제공한다.

위험관리 효과는 이 프로세스를 통해 책임 있는 AI 실행, 개발 및 사용을 추적하고 보장할 수 있다는 것이다.

AI 시스템 수명주기를 분석을 위한 지침 프레임워크로 사용하는 이 툴킷은 AI 기술을 사용하여 의사결정 프로세스와 결과를 개선하려는 공공 정책 팀에 기술 지침을 제공한다.

AI 시스템 수명주기의 각 단계(계획 및 설계, 데이터 수집 및 처리, 모델 구축 및 검증, 배포 및 모니터링)에 대해 이 툴킷은 공공 정책 맥락에서 AI를 사용할 때 발생하는 가장 일반적인 과제를 식별하고 탐지 및 완화를 위한 절차를 수행하고 있다[13][14][15][16].

라. RAI 툴킷의 위험관리 평가 원칙 및 특징

RAI 툴킷은 6가지 순차적 활동의 약어인 SHIELD (Set foundation, Hone operationalization, Improve & Innovate, Evaluate Progress, Log for traceability, Detect via continuous monitoring) 평가를 기반으로 구축하고 있다.

Shield process 구성 요소인 6가지 순차적 활동을 살펴보면, (1) 기본 설정(Set foundation), (2) 운영화(Hone operationalization), (3) 개선과 혁신(Improve & Innovate), (4) 상태 평가(Evaluate Progress), (5) 추적성을 위한 로그(Log for traceability), (6) 지속적인 모니터링을 통한 탐지 (Detect via continuous monitoring) 이다.

표 3. RAI 툴킷의 5원칙
Table 3. Five Principles of the RAI Toolkit

구분	내용
모듈식 및 맞춤형	RAI 툴킷은 모든 프로젝트가 사용 사례, 컨텍스트 또는 우선순위가 다르기 때문에 맞춤형으로 제작함 툴킷 사용자는 관련 레이블을 태그하여 표시되는 콘텐츠를 맞춤화할 수 있으며, 관련 없는 콘텐츠를 제외하고 꼭 필요한 것만 볼 수 있음
RASCI 매트릭스에 맞춰 조정	RAI 툴킷은 RASCI(Responsible:책임, Accountable:책임자, Support:지원, Consulted:상담, Informed:정보제공) 매트릭스를 사용함 책임성을 높이는 동시에 전반적인 프로세스를 더욱 효과적으로 제공함 RASCI 매트릭스에 대한 이러한 정렬은 프로그램 관리자(PM)가 프로그램에 대한 완전한 지식을 가지고 있다고 가정하지 않음
주요 내용	RAI 툴킷은 AI 시스템을 평가하고 개선하기 위한 리소스를 제공함 단순히 엔지니어링에 관한 것이 아니며 툴킷은 책임감 있는 개발 및 사용 문화를 조성함 기술 거버넌스부터 시작하지만 향후 버전에는 조직 거버넌스 및 운영 지침을 위한 리소스가 포함 예정
윤리 원칙	RAI 툴킷에는 DoD AI 윤리 원칙이 포함되어 있으며 사용자는 각 요소를 해당 원칙으로 추적할 수 있음
도구 목록	RAI 도구 키트는 위험을 완화하거나 AI 시스템 개발을 개선하는 데 도움이 되는 70개 이상의 도구 목록을 제공함 툴킷에서는 대부분의 도구가 업계 표준 오픈 소스 옵션임

RAI 툴킷은 다음과 같은 특징을 제공한다.

첫째, 기초설정이다. 기초 설정 잠재적인 위험, 피해, 기회 및 영향과 함께 프로젝트에 대한 관련 RAI, 윤리적, 법적 및 정책적 기초를 식별한다. 그리고 제품 수명주기 전반에 걸쳐 추적할 수 있는 문제 목록을 만든다.

둘째, 운영 개선 원칙이다. 운영 개선 원칙이 어느 정도 충족되고 문제가 해결되고 있는지 평가하는 방법을 결정하기 위해 기초와 SOC를 구체적인 방법으로 운영한다.

셋째, 개선 및 혁신 완화 도구의 활용이다. 개선 및 혁신 완화 도구와 활동을 활용하여 기초를 충족하고 SOC를 해결하는 과정을 개선한다. 그리고 최소 요구 사항 이상으로 기술을 더욱 향상하기 위한 새로운 혁신의 범위를 정하고 구현한다.

넷째, 평가과정이다. 진행 상황 벤치마크를 평가하고 기반이 어느 정도 충족되고 있는지, SOC가 해결되고 있는지, 모든 혁신이 기준선에서 개선되고 있는지 평가한다.

다섯째, 추적관리이다. 추적성을 위한 기록 확인 문서는 추적성을 보장하고 학습한 교훈을 저장소에 공유하기 위한 것이다. 새로운 평가 또는 완화 방법이 필요한 위치를 추적한다.

여섯째, 모니터링이다. 지속적인 모니터링을 통해 시스템 성능 저하를 감시한다.

마. RAI 수행 절차

RAI 툴킷은 다음과 같은 7가지 수행 절차를 통해 수행하고 있으며 단계별 상세내용이다.

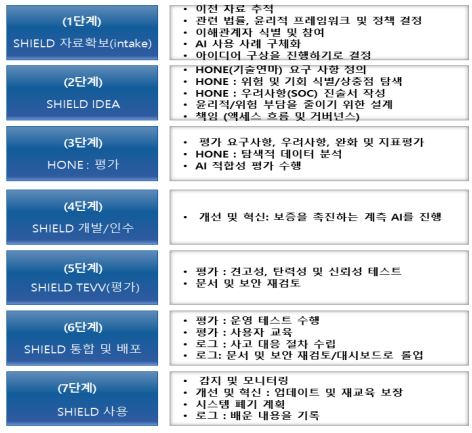

그림 3. RAI 작업 순서도
Fig. 3. RAI operation flow chart

1단계인 자료 확보단계에서 다음과 같은 내용이 진행된다. 첫째. 이전 자료추적에서 현재 프로젝트에 적용할 수 있는 RAI 관련 "교훈"을 식별하기 위해 사고 저장소를 포함하도록 사용 사례와 비교하여 유사한 프로젝트를 검토한다. 그리고 검토 내용을 RAI 사용 사례 저장소, AI 사고 저장소, AI 사고 데이터베이스에 저장한다. 검토를 통해 나온 관련 법적, 윤리적, 정책적 지침과 고려 사항을 문서화 관련 거버넌스 요구 사항은 모든 DoD 법적 및 정책 기반 요구 사항이 포함된 포괄적인 저장소를 확인한다.

둘째, 관련 법적, 윤리적, 위험 및 정책 프레임워크를 식별하고 해당 프레임워크의 적용 가능성을 평가한다. 관련 법률, 규정, 정책 및 프레임워크(법적, 윤리적, 위험 등) 중 어느 것이 프로젝트에 적용되는지 결정하고 이러한 적용 가능성 결정을 뒷받침하는 근거를 문서화한다.

AI 프로젝트의 다운스트림 문서가 결정된 법률, 규정, 정책 및 프레임워크에 해당하는 산출물 및 표준과 일치하는지 확인한다. 또한, 조직 전체의 정책, 프로세스, 절차 및 관행이 이러한 법률, 규정, 정책 및 프레임워크에 매핑되어 있으며 투명하고 효과적으로 구현되었는가? 적용 가능한 각 프레임워크에 대한 권위 있는 참고 자료가 있고 이에 대해 상담할 수 있는 해당 분야 전문가가 식별되어 있는가? 법적 검토 계획이 있는가? 애플리케이션 도메인이 변경되거나 정기적으로 리뷰를 업데이트하는 절차를 확립했는가? AI 프로젝트의 윤리적, 법적, 안전 위험을 담당할 책임이 있는 적절한 고위 책임자를 식별했는가? 등에 대해 자료 수집을 한다.

2단계에서 SHIELD IDEA단계로서 (1) HONE(기초) 요구 사항 정의, (2) HONE : 위험 및 기회 식별/상충점 탐색, (3) HONE : 우려사항(SOC) 진술서 작성 , (4) 윤리적/위험 부담을 줄이기 위한 설계, (5) 책임 (액세스 흐름 및 거버넌스)를 수행한다.

3단계는 HONE 평가 단계로서 (1) 평가 요구사항, 우려사항, 완화 및 지표평가, (2) HONE : 탐색적 데이터 분석, (3) AI 적합성 평가 수행한다.

4단계는 SHIELD 개발/인수 단계로서 (1) 개선 및 혁신: 보증을 촉진하는 계측 AI를 진행한다.

5단계인 SHIELD TEVV(Test System for Robustness, Resilience, and Reliability, 평가)에서는 (1) 평가 : 견고성, 탄력성 및 신뢰성 테스트, (2) 문서 및 보안 재검토 한다.

6단계인 SHIELD 통합 및 배포에서는 (1) 평가측면에서 운영 테스트 수행, (2) 사용자 교육 평가, (3) 로그 기반 사고 대응 절차 수립, (4) 문서 및 보안 재검토/대시보드로 롤업을 수행한다.

7단계인 SHIELD 사용에서는 (1) 감지 및 모니터링, (2) 개선 및 혁신 : 업데이트 및 재교육 보장, (3) 시스템 폐기 계획, (4) 로그 영역은 배운 내용을 기록을 수행한다.

IV. 결 론

금융, 보험, 농업, 교통에 이르기까지 인공지능(AI) 기술은 모든 분야에서 빠르게 확산되어 기회를 창출하는 동시에 이로 인해 등장하는 위험을 어떻게 관리 할 것인가에 대해 관심이 쏠리고 있다.

본 연구의 목적은 인공지능과 위험관리라는 2가지 키워드를 통해 어떻게 인공지능 서비스의 장점 활용과 위험요인을 동시에 극복하는데 기여 하고자 한다. 이를 위해 2가지 사례인 (1) 인공지능을 활용한 위험 모니터링 프로세스 제시와 (2) 인공지능 서비스의 개발 및 운영에서 등장하는 한계 요인을 최소화하기 위한 운영 툴킷에 대해 소개하였다.

표 4. 인공지능과 위험관리 사례
Table 4. Artificial intelligence and risk management cases

구분	진행 프로세스
인공지능 활용 위험모니터링(AIM) 과정	(1단계) 딥러닝 모델 사용 (2단계) 프레임워크로 분류 및 분석 수행 (3단계) 평판 우수 데이터 확보 (4단계) 내용 수집 및 분석 (5단계) 분석 결과 링크 제공
RAI 툴킷의 적용 프로세스	(1단계) SHIELD 자료확보(intake) (2단계) SHIELD IDEA (3단계) HONE : 평가 (4단계) SHIELD 개발/인수 (5단계) SHIELD TEVV(평가) (6단계) SHIELD 통합 및 배포 (7단계) SHIELD 사용

첫째, 웹사이트에서 등장하는 내용을 기반으로 위험 내용을 구체적으로 추적하는 절차로서 인공지능을 활용한 위험 모니터링에 대해 살펴보았다.

이를 수행하기 위해서는 적절한 딥러닝 프로세스가 존재해야 하며 적절한 프레임 분류체계와 가짜 뉴스가 아닌 평판이 우수한 데이터 확보가 이루어져야 한다. 그리고 제작된 데이터에 대한 검증과 오류보정이 이루어져야 한다. 이 과정에 대한 충분한 이해를 통해 우리가 접하는 데이터의 신뢰성에 대한 검토가 이루어져야 한다.

이는 기본적으로 데이터 수집과 분류 등 결과를 인공지능 기술로 해결하고 있지만, 여전히 많은 영역이 인간이 개입하여 조정하고 관리해야 부분이 존재한다. 그럼에도 불구하고, 인공지능 기술은 기존의 많은 시간과 노력이 요구되는 영역을 신속하게 효율적으로 해결하고 있다는 것에 많은 의의를 가진다. 상기 사례는 위험관리 모니터링에 대한 일반적인 절차를 인식하고 이를 지속적으로 관리하여 위험관리 모니터링 서비스에 적용하는 것에 대한 방법론을 제공하는 것에 시사점을 가진다. 또한, 이를 위해 가장 중요한 요소가 우수하고 적절한 데이터 확보가 중요하며 활용되는 데이터에 신뢰가 우선적으로 이루어져야 한다.

두번째 제공된 사례인 인공지능 서비스 위험관리 RAI 평가 도구는 AI 수명주기의 복잡한 단계를 통해 개인을 위한 리소스 역할을 하여 프로젝트 내에서 RAI가 핵심적인 차원을 탐색할 수 있도록 지원한다. 이는 유사한 노력에서 얻은 모범적인 결과물 및 통찰력과 함께 해당 부서가 지원하는 다양한 도구를 제공함으로써 업무 효율성을 지원한다.

미국 국방부내 RAI 구현의 일환으로 이 툴킷은 사용자에게 미국국방부의 AI 기능을 책임감 있게 개발, 배포 및 사용하는 데 도움이 되는 다양한 고려 사항을 제공하는 것을 목표로 한다.

미국 국방부의 책임 있는 인공지능 전략 및 구현 경로에 명시된 바와 같이, 국방부는 AI를 설계, 개발, 테스트, 조달, 배치 및 사용할 때 합법적이고 윤리적인 행동에 대한 우리 군대의 확고한 의지가 적용된다는 것을 입증해야 한다.

본 연구는 다음과 같은 시사점이 제안하고자 한다. 첫째, 인공지능 서비스는 우리 삶에 깊숙이 관여를 하고 있으며 이로 인해 등장하는 한계요인을 최소화 하는 장치가 필요하다. 둘째, 인공지능을 활용한 위험관리 모니터링은 적합하고 신뢰성이 있는 데이터 확보가 우선적으로 고려되어야 한다. 셋째, 인공지능 서비스의 개발과 운영 시 등장하는 한계를 극복하기 위해서는 업무 단계별로 위험관리 프로세스를 적용하여 상시 모니터링이 요구된다는 것이다.

본 연구를 통해 등장하고 인공지능을 활용하여 위험관리를 탐색하는 것과 인공지능 서비스의 위험관리를 하는 방안에 대해 사례 분석을 통해 우리가 인공지능과 위험관리를 어떻게 접근해야 할지에 대한 시사점을 제공한다. 이를 통해 인공지능 서비스의 위험관리를 어떻게 바라보고 접근할 수 있는 것에 대한 정책 수립에 기여하고 있다.

본 연구는 인공지능과 위험관리라는 키워드를 중심으로 2가지 사례를 분석한 것으로 관련 분야에 대한 다양한 연구가 지속적으로 이루어져야 한다. 이는 등장하는 서비스에서 인공지능 기술의 장점을 획득하고 이로 인해 등장하는 단점 요소를 최소화 하는게 현재의 우리의 남겨진 숙제이다.

우리는 본 연구를 통해서 살펴본 것처럼 인공지능에 제공하는 장점을 수용하면 한계 및 위험요소에 대한 적극적인 대응 방안으로 마련해야 한다. 인공지능 기술을 적극적 활용하여 효율적으로 정책 수립하여 고부가가치를 생성하고 사회에 의미 있는 영향을 제공할 수 있도록 노력해야 한다.

본 연구는 발전하고 있는 인공지능이 제공하고 한계요인을 최소화 할 수 있는 방안에 대한 연구로서 향후 이와 관련 시장의 성장과 발달에서 위험관리 측면에 대한 연구에 기여한다. 아울러 인공지능과 위험관리에 대한 2가지 사례 측면에서 논의한 것이 한계이며 향후 연구는 다양한 측면에서 보다 더 구체적으로 논의 및 연구가 이루어져야 한다.

References

[1] SAS Website, https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html

[2] Ryu, Young-ho. "Artificial Intelligence and Changes in the Publishing Industry "https://nzine.kpipa.or.kr/sub/zoomin.php?ptype=view&idx=423&page=&code=zoomin&total_searchkey=%EC%B2%B4%ED%97%9

[3] Yoo, Soonduck. "Research on the evaluation model for the impact of AI services." International Journal of Internet, Broadcasting, and Communication 15, no. 3, pp 191-202, 2023. DOI http://10.7236/IJIBC.2023.15.3.191

[4] Microsoft OneNote Privacy Policy for Android - Microsoft Support

[5] National Intelligence Informationization White Paper 2022, 2022

[6] https://spacy.io/

[7] https://wikifier.org/

[8] Shin, Sungpil. "Concept and Standardization Trends of Foundation Models for Ultra-Large AI." Journal of the Korean Institute of Communication Sciences (Information and Communication) Vol 40, No. 6, pp 12-21, 2023.

[9] Kenton, Jacob Devlin, Ming-Wei Chang, and Lee Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." In Proceedings of NAACL-HLT, Vol. 1, PP 2-10. 2019.

[10] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized BERT pretraining approach." arXiv preprint arXiv:1907.11692, 2019.

[11] Li, Bangzheng, Wenpeng Yin, and Muhao Chen. "Ultra-fine entity typing with indirect supervision from natural language inference." Transactions of the Association for Computational Linguistics Vol 10, pp 607-622, 2022.

[12] Soares, Livio Baldini, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. "Matching the blanks: Distributional similarity for relation learning." arXiv preprint arXiv:1906.03158, 2019.

[13] Lee, Deok-hee. "Establishing the Infrastructure of a Leading Innovative Country in the Fourth Industrial Revolution."

[14] Artificial Intelligence in Society PDF, Korea Artificial Intelligence Society for Promotion, 2019

[15] Choi, Nak-Hun, et al. "A development of defeat prediction model using machine learning in polyurethane foaming process for automotive seat." Journal of the Korea Academia-Industrial cooperation Society, Vol 22. NO. 6, pp 36-42, 2021
DOI : http://10.5762/KAIS.2021.22.6.36

[16] Sohee Par, et. al. ""Risk Assessment of Actuators Uncertainty using STPA and SMC" Journal of KIIT. Vol. 21, No. 8, pp. 39-49, 2023
DOI http://dx.doi.org/10.14801/jkiit.2023.21.8.39

저 자 소 개

신 선 영(정회원)

- 동국대학교 컴퓨터공학사
- 연세대학교 산업공학 석사
- 경북대학교 경영정보학 박사
- 2001년 ~ 현재 : 한국지능정보사회진흥원(NIA) AI데이터 활용팀장
- 관심분야 : AI 사회적 영향평가, AI 기반 의사결정, 빅데이터 분석, ICT 정책 수립, AI데이터