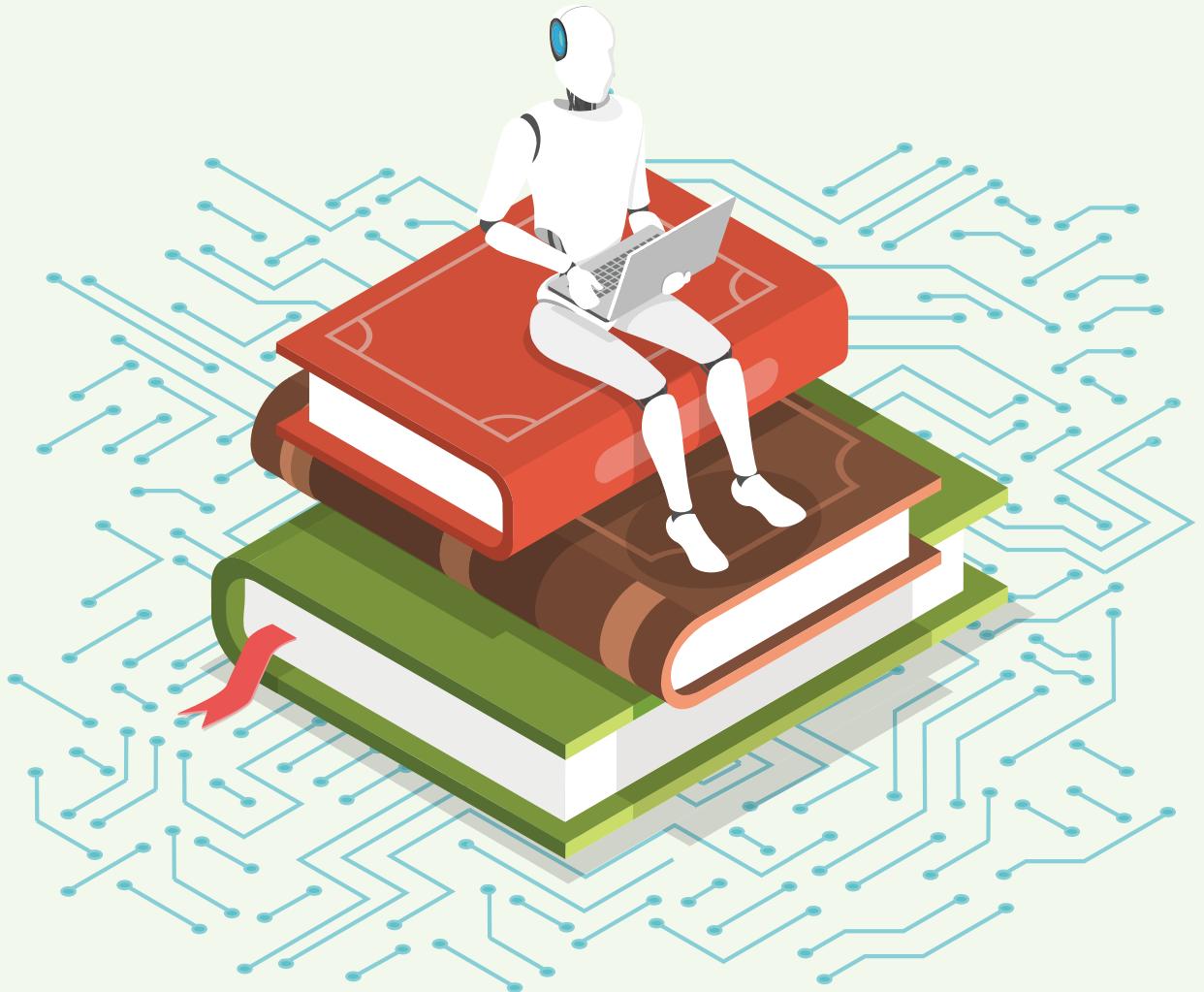


2022 신뢰할 수 있는 인공지능 개발 안내서

신예진 TTA AI시험검증팀 선임연구원



1. 머리말

인공지능(AI, Artificial Intelligence) 기술은 우리의 삶과 밀접한 의료, 법률, 공공안전 등 다양한 산업에 도입되면서 지속적인 발전을 이루고 있다. 인공지능은 질병을 진단하고, 고객을 응대하고, 각종 위기 상황을 사람보다 빠르게 인지하는 등 인간의 조력자이자 동반자가 될 기술로 손꼽힌다.

하지만 그 이면으로, 인공지능 활용 과정에서 위험·부작용 등이 발생하기도 한다. 2020년 10월, 한 영국 매체에서 프랑스 헬스케어 기업인 Nabla가 개발한 인공지능 정신과 상담 챗봇을 시범 운영한 결과, 모의 환자에게 자살 충동을 실행에 옮길 것을 권유하는 사례가 있었다. 이렇듯, 인공지능의 확산과 함께 우리의 안전이나 재산 등에 직·간접적인 피해를 줄 우려도 나타나고 있다.

이를 해결하고자 주요 선진국 및 국제기구에서는 인공지능 관련 가이드라인(UNESCO 윤리권고[1], OECD AI Principle[2], EU GDPR[3] 등)을 내놓고 있고, 국제표준화기구 ISO/IEC의 인공지능 위원회 JTC1/SC42에서는 인공지능 관련 기술 영역의 표준화 작업을 주도하고 있다. 특히, SC42 산하 그룹 중 신뢰성 작업그룹(WG3)은 투명성(Transparency), 설명가능성(Explainability) 등 인공지능 신뢰성 내용의 표준화를 논의하고 있다. 이런 흐름에 발맞춰 국내에서도 과학기술정보통신부가 지난해 5월(사람이 중심이 되는 인공지능을 위한 「신뢰할 수 있는 인공지능(AI) 실현 전략」[4])을 발표했다.

이처럼 인공지능의 개발 및 활용에 대한 윤리 및 신뢰 측면의 권고나 규제가 마련되고 있지만 내용이 거시적이고 추상적인 관점에서 작성되어

있어, 해당 내용을 인공지능 기술에 알맞게 적용하고 평가하는 방법론 및 기준은 아직 모호하다. 따라서 기술적, 공학적으로 인공지능 제품·서비스를 설계하거나 개발하는 실무자가 참고할 만한 자료가 필요한 실정이다.

과학기술정보통신부와 TTA는 이러한 필요성에 공감하여 인공지능의 신뢰성 확보를 위한 기술적 수단을 고민해왔고, 그 결과로 '2022 신뢰할 수 있는 인공지능 개발 안내서[5]'(이하 안내서)를 공개하였다. 본고에서는 안내서의 개발 과정과 안내서에서 제시한 인공지능 신뢰성 자가점검을 위한 요구사항 및 검증항목의 활용 방안을 기술하고, 향후 안내서의 발전 방향을 소개한다.

2. '2022 신뢰할 수 있는 인공지능 개발 안내서' 개발 과정

2.1 인공지능 신뢰성의 중요성

ISO/IEC TR 24028 Overview of trustworthiness in artificial intelligence[6]에서는 신뢰성(Trustworthiness)을 '검증 가능한 방식으로 이해관계자의 기대에 부응하는 능력'이라고 정의했다. 또한 EC의 Ethics Guidelines for Trustworthy AI[7]에서는 신뢰할 수 있는 인공지능(TAI, Trustworthy AI)은 관련 법규를 준수하고, 윤리적 원칙과 가치를 준수하며, 좋은 의도에도 불구하고 의도치 않게 발생할 수 있는 피해에 대해 기술적으로나 사회적으로 강건하다는 것을 의미한다고 서술했다. 이렇듯 인공지능 신뢰성은 인공지능의 위험과 부작용을 선제적으로 파악하고, 기술·사회·윤리·제도적 측면에서 종합적인 대비 및 검증에 필수적인 요소로 인식할 필요가 있다.

2.2 인공지능 신뢰성 확보에 필요한 요소 탐색

머리말에서 밝힌 바와 같이, 그간 국내외 많은 기관 및 기업이 인공지능 신뢰성 확보를 위해 윤리 원칙과 지침·가이드라인을 내놓았으나, 기술적 관점에서 상세한 방법론을 제시한 사례는 없었다. 따라서 TTA에서는 인공지능 제품 및 서비스 개발 현장의 데이터 과학자, 모델 개발자 등 이해관계자들이 실무 관점에서 신뢰성 확보에 참고할 수 있는 지침서 성격의 자료를 만들고자 했다. 이러한 목적에 따라 안내서 개발 과정에서 신뢰성 확보를 위해서는 어떤 요소들이 실무적으로 고려되어야 하는지 탐색하였고, 그 결과 세 가지 요소를 반영하였다.

첫 번째는 **인공지능 서비스의 구성 요소**이다. 인공지능 서비스의 핵심적인 사고 기능을 수행하는 인공지능 모델 및 알고리즘, 이를 학습시킬 데이터, 실제 기능이 구현될 소프트웨어 기반 시스템, 사용자와 상호작용하기 위한 인터페이스가 구성 요소에 해당한다. 이 구성 요소들은 개별적으로, 또는 통합되어 인공지능 서비스의 생명주기에 따라 개발·검증·운영된다. 따라서 구성 요소별 신뢰성 확보 방안을 고민하고, 각 요소에 따른 요구사항과 검증항목을 제시하고자 했다.

두 번째는 **인공지능 서비스의 생명주기**인데, 이것은 첫 번째에서 살펴본 구성 요소들을 구현하고 운영하는 과정에 해당한다. 기존 소프트웨어 시스템의 생명주기와 유사하나, 인공지능 기술에 맞춰 데이터 처리 및 모델 개발 단계가 추가된 점이 특징이다. 각종 문헌에서는 인공지능 서비스 생명주기를 6~8단계로 구분하여 정의하지만, 안내서에서는 실무자들이 쉽게 활용할 수 있도록 각 단계의 성격과 활동을 왜곡하지 않는 선에서 5가지 단계로 단순화했다. 안내서에서 정의한 5가지 생명주기는 ‘계획 및 설계’, ‘데이터

수집 및 처리’, ‘인공지능 모델 개발’, ‘시스템 구현’, ‘운영 및 모니터링’ 단계로 이루어진다. 물론, 인공지능 서비스의 특성상 생명주기의 일부 단계들은 반복, 순환될 수 있다.

마지막으로, 세 번째는 ‘인공지능(AI) 윤리기준[8]’의 핵심 요건을 준용한 **기술적 관점의 신뢰성 요건**이다. 많은 문헌에서는 인공지능 신뢰성을 하위 속성으로 세분화해 제시하고 있다. 이 속성을 종합적으로 분석해본 결과, 관점에 따라 유사해 보이나 조금씩 다른 용어들이 제각기 정의되고 있었다. 따라서 델파이(Delphi) 기법 등을 활용하여 국내 각계 전문가들의 의견을 수렴하고 합의점을 모색한 후, 이를 인공지능(AI) 윤리기준의 10대 요건에 대응시켰다. 그 결과, 기술적 측면에서 고려되어야 할 요건으로 ‘다양성 존중’, ‘책임성’, ‘안전성’, ‘투명성’이 최종 선정되었다.

2.3 인공지능 신뢰성 확보를 위한 요구사항 도출

2.2절에서 소개한 인공지능 서비스 구성, 생명주기, 신뢰성 요건을 토대로 구체적인 요구사항과 검증항목을 도출했다. 우선 국제기구, 기술단체, 표준화기구 및 주요 국가 정부에서 발표한 인공지능 신뢰성 확보 정책, 권고안, 표준 등[6,8-16]을 기반으로 준수해야 할 기술적 요구사항을 도출하고 구체화하였다. 이 과정에는 금융·분야 AI 가이드라인[17] 등 국내에서 발표된 사례들도 포함되었다.

다양한 문헌들을 기반으로 폭넓은 범위의 기술적 요구사항들을 도출하고 중복 내용을 제거하는 등 TTA 연구진의 면밀한 검토 과정을 거쳤다. 또한 선별된 요구사항들의 만족 여부를 확인하기 위한 검증항목을 마련하였다. 이 과정에서 각 검증항목이 요구사항에 잘 부합하는지, 개발 현장에서 실무적으로 활용 가능한지, 요구사항

<표 1> 인공지능 신뢰성 확보를 위한 기술적 요구사항과 신뢰성 요건

요구사항	다양성 존중	책임성	안전성	투명성
01. 인공 지능 시스템에 대한 위험관리 계획 및 수행		○		○
02. 데이터의 활용을 위한 상세 정보 제공		○		○
03. 데이터 강건성 확보를 위한 이상(Abnormal) 데이터 점검			○	
04. 수집 및 가공된 학습 데이터의 편향 제거	○	○		○
05. 오픈소스 라이브러리의 보안성 및 호환성 확보		○	○	
06. 인공지능 모델의 편향 제거	○			
07. 인공지능 모델 공격에 대한 방어 대책 수립			○	
08. 인공지능 모델 명세 및 출력 결과에 대한 설명 제공		○		○
09. 인공지능 모델 출력에 대한 신뢰도(Confidence Value) 제공				○
10. 인공 지능 시스템 구현 시 발생 가능한 편향 제거	○			
11. 인공 지능 시스템의 안전 모드 구현		○	○	○
12. 인공 지능 시스템의 설명에 대한 사용자의 이해도 제고				○
13. 인공 지능 시스템의 추적 가능성 확보			○	○
14. 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공		○		○

<표 2> 인공지능 생명주기 단계별 요구사항의 확인 및 검증 주체

생명주기 단계	해당 요구사항	확인 및 검증 주체
1. 계획 및 설계	요구사항 01	- 비즈니스 결정권자 - 시스템 기획자 - 데이터 과학자 - 시스템 운영자
2. 데이터 수집 및 처리	요구사항 02-04	- 데이터 공급자 - 데이터 과학자 - 도메인 전문가
3. 인공지능 모델 개발	요구사항 05-09	- 데이터 과학자 - 인공지능 모델 개발자 - 시스템 엔지니어
4. 시스템 구현	요구사항 10-12	- 인공지능 모델 개발자 - 시스템 엔지니어
5. 운영 및 모니터링	요구사항 13-14	- 비즈니스 결정권자 - 인공지능 모델 개발자 - 시스템 운영자

과 검증항목이 인공지능 기술에 대한 연구내용을 폭넓게 포함하는지 등 기술적 타당성과 효용성, 포괄성 측면에서 확인을 거듭하였다. 이를 위해 기획자, 개발 프로젝트 리더, 교수, 연구원, 정책담당자 등 다수의 인공지능 분야 전문가가 참여하여 검토하고 자문했으며, 다양한 의견들을

수렴하고 반영했다.

최종적으로, 자율적으로 검증 가능한 14개 요구사항 및 이에 매칭되는 59개 정성·정량적 검증 항목이 도출되었다. 14개 요구사항의 목록과 각 요구사항에 해당하는 신뢰성 요건을 표시하여 <표 1>에 정리하였다.

3. 인공지능 신뢰성 자가점검을 위한 요구사항 및 검증항목 활용 방안 안내

3.1 요구사항 확인 및 검증 주체

안내서는 인공지능 서비스를 구현하는 과정에 직·간접적으로 관련되거나 영향을 주는 모든 조직과 개인을 포함한 이해관계자를 대상으로 제작되었다. 특히 업무상 기술적 관점에서 신뢰성을 신경 써야 하는 기획자, 데이터 수집 및 가공자, 인공지능 모델 개발자, 시스템 및 소프트웨어 개발자, 테스터 등이 주요 대상이다. 이들은 인공지능 생명주기의 각 단계마다 인공지능 신뢰성을 확보하기 위해 요구사항을 확인하고 검증하는 주체로 활동해야 할 것이다.

인공지능 생명주기 단계별로 요구사항 확인 및 검증 주체를 대응시킨 결과를 <표 2>에 나타내었다. 물론, <표 2>에 제시한 대상이 요구사항 확인 및 검증 시 주도적으로 활동하되, 이외 다양한 이해관계자와의 협력을 통해 활용되는 것이 바람직하다. 주체와 협력 대상은 안내서를 활용하는 서비스 및 기업 환경에 따라 상이할 수 있으므로, <표 2>의 내용은 권장 사항으로써 활용되길 바란다.

3.2 요구사항 확인 및 검증 절차

안내서를 활용한 요구사항 확인 및 검증 절차는 크게 3단계로 정리할 수 있다. 첫 번째는 **인공지능 서비스 유형을 검토하는 단계**이다. 이 단계에서는 인공지능 서비스의 활용 목적과 범위, 활용 대상, 문제 발생 시 영향도를 분석한다. 그리고 인공지능의 예측 결과에 대해 최종적으로 사람이 판단하는지, 또는 인공지능의 예측 결과를 사람이 그대로 받아들이는지 등 사람의 개입 여부와 개입 정도를 검토한다.

두 번째 단계에서는 확인 및 검증해야 할 요구사항을 선별한다. 첫 번째 단계에서 검토한 결과를 토대로 대상 서비스에 적용되어야 할 요구사항을 선정하는 과정이다. 안내서 내 요구사항 요약문과 세부 요구사항 본문을 참고하여, 서비스에 필요한 요구사항인지 여부를 판단할 수 있다. 이때, 서비스에서 필요하지 않을 것으로 판단한 요구사항이 있다면 안내서 내 체크리스트에서 'Y/N/NA' 중 'NA'로 표시하여 검증 대상에서 제외한다.

마지막 단계로, **선별된 요구사항에 대한 검증을 수행한다**. 이때, 세부 요구사항 및 검증항목을 참고하여, 요구사항별 만족 여부를 확인하는 과정이다. 검증항목에 따라 충족 여부가 정성적으로 평가될 수 있으나, 이는 검증 주체 및 협력 대상자들이 협의하여 첫 번째 단계에서 검토한 서비스 유형과 영향도에 따라 판단하도록 한다. 검증항목의 평가 결과 해당 항목이 충족되었음을 확인한 경우, 안내서 내 체크리스트에서 'Y'로 표시한다. 반대로, 검증항목의 평가 결과 해당 항목이 미충족되었다면, 안내서 내 체크리스트에서 'N'으로 표시한다.

물론 인공 지능 시스템은 운영 과정에서 그 형태와 형상이 변화할 수 있는 특성이 있어, 시스템에서 새로운 위험요소가 지속적으로 발생할 가능성이 있다. 따라서 요구사항 확인 및 검증 활동은 일회성으로 끝나지 않고 생명주기 전반에 걸쳐 반복적으로 이루어져야만 적절한 위험관리가 이루어질 수 있다. 더불어, 안내서에서 소개된 기술적 요구사항 외에도 인공지능(AI) 윤리 자율점검표[18], 인공지능(AI) 개인정보보호 자율점검표[19]를 참고하고, 조직 관점의 가이드라인이나 전통적인 소프트웨어 시스템의 성능, 보안 등 품질 관점의 표준과 같은 다양한 방

법률의 적용을 검토하는 것이 바람직하다.

4. 향후 '신뢰할 수 있는 인공지능 개발 안내서' 발전 방향

4.1 학계·산업계 반응

안내서 개발 과정에서 학계·연구계·산업계 전문가 자문을 통해 안내서에 대한 검토 및 보완을 지속해서 수행하였으며, 안내서 공개 후에도 과학기술정보통신부에서 출범한 '인공지능 윤리 정책 포럼' 제2분과 위원들을 대상으로 의견을 수렴하는 등 총 12회에 걸쳐 250여 명의 전문가와 실무자 대상 의견수렴을 진행하였다.

전문가들은 인공지능 개발 실무자가 참고 가능한 자가 점검항목을 마련한다는 목적과 필요성·방향성에 공감하며, 동시에 다음과 같은 긍정적인 의견들을 제시하였다.

- ① 안내서를 통해 각 기업에서 신뢰성 확보를 위한 방향성 설정이 가능할 것으로 기대
- ② 인공지능 구현 생명주기에 따라 고려해야 하는 요구사항 및 검증항목을 제시한 안내서 구성 방식이 활용하기에 적절
- ③ 가독성 및 이해가능성이 높아 초급 실무자가 이해할 수 있는 수준으로 접근성 확보

반면에 안내서의 아쉬운 부분과 발전 방향을 제시하는 다음과 같은 목소리도 들을 수 있었는데, 실제 산업 현장에 실효성 있는 안을 도출하기 위한 지속적인 노력이 필요할 것이다.

- ① 요구사항 적용 사례나 검증 예시가 추가된다면 실제 현장에서 바로 활용하는 데 더 도움이 될 것으로 예상
- ② 일부 분야에서는 적용이 불필요한 요구사항도 있고, 동시에 특정 분야에서는 깊이 있는 검증이 필요한 요구사

항도 있어 영역 및 분야에 따른 구분이 필요

③ 신뢰성 확보를 위해 필요하지만, 현재는 기술이 발전하는 단계에 있어 일부 요구사항은 바로 적용되기에 한계가 있거나, 초기대 모델 등 모든 데이터를 검증하기에 현실적으로 어려움이 있는 요구사항이 존재

4.2 현장 중심의 안내서 마련을 위한 보완 계획

4.1절에서 살펴본 학계·산업계 의견을 바탕으로, 시급성 및 타당성을 고려하여 다음의 3가지 측면에서 안내서 보완 작업을 진행하는 중이다. 보완 과정에서 현장과 소통하면서 안내서가 산업 현장에서 활용될 수 있도록 지속적인 개선을 이어나가고자 한다.

① 특정 분야에 테일러링된 안내서 추가 마련

특정 영역 또는 분야를 중심으로 신뢰성 확보를 위한 수단을 검토해야 한다는 측면의 전문가 의견에 공감하며, 앞으로는 인공지능 기술을 활용한 다양한 서비스와 제품 영역별로 안내서를 고도화해나갈 계획이다. 사전 검토·분석을 통해 인공지능 도입 및 활용이 활성화된 산업 분야 3개를 선정하였고, 올해에는 의료, 자율주행, 공공·사회 분야의 안내서를 마련하고자 한다. 분야별 안내서는 기존 안내서를 기반으로 분야별 특성에 따라 다르게 적용될 수 있는 세부 요구사항 및 검증항목에 대한 설명을 추가할 예정이다. 이를 활용하여 실제 서비스를 개발하고 운영하는 현장에 적용해본 후, 요구사항 및 검증항목 적용사례와 해당 분야에서 활용되는 데이터 및 모델에 적용 가능한 도구·기법 등의 안내를 통해 실무자의 이해도를 제고하고자 한다.

② 범용성을 갖춘 형태의 기존 안내서 고도화

기존 안내서는 신뢰할 수 있는 인공지능 구현을 위한 큰 흐름을 일목요연하게 정리하고 요구사항과 검증항목들이 신뢰성 전체 영역을 포괄적으로 보여줄 수 있도록 제시한 상태로, 계속해서 범용성을 갖춘 형태로 고도화를

진행할 계획이다. 고도화 진행 방향은 기업·기관 대상 적용 사례와 피드백을 종합하여 범용적인 관점에서의 개선 사항을 반영하고 수록하며, 국내외 정책 동향과 최신 기술을 반영하는 것에 초점을 맞출 예정이다.

- ③ 전문가 의견수렴 및 사례 기반의 현실적 검증방안 도출 요구사항과 검증항목의 현실적·기술적 적용 가능성에 대한 다양한 의견과 기술적 한계를 인정하고, 안내서에서 단서조항의 성격으로 추가 기술할 계획이다. 대표적인 예시로 설명가능성이 있는데, 설명가능성은 데이터나 모델에 따라 구현이 어려울 수 있고, 설명가능성의 필요성이나 그 정도에 대해 각계 의견이 분분하다. 따라서 모든 인공지능 제품에 설명가능성을 적용하는 것이 아니라 제품 및 서비스의 다양성에 대한 고려나 설명가능성이 미치는 영향에 대한 고려가 선행되어야 할 것이다. 이를 위해 현장 적용 및 컨설팅 등을 수행하여 사례 기반으로 상세 내용을 도출할 예정이다.

5. 맷음말

본고에서는 '2022 신뢰할 수 있는 인공지능 개발 안내서'의 개발 과정과 안내서에서 제시한 인공지능 신뢰성 측면 요구사항 및 검증항목의 활용 방안, 향후 안내서의 발전 방향에 대해 살펴보았다. 인공지능 신뢰성은 연구계, 산업체를 막론하고 어느 분야에서나 중요한 주제로 연구와

논의가 지속되고 있는 분야이며, 사회 구성원의 다양한 의견과 논의를 통해 합의와 공감대를 이루어야 하는 개념이다.

이를 위해, 인공지능 윤리나 신뢰성을 제품에 구현하는 기술적 방안에 대한 고민이 필요한 시점이다. 그러나 아직까지 인공지능 신뢰성과 관련된 요구사항이나 검증방안이 자세하게 표준이나 가이드라인 등으로 문서화된 바는 없어 실무자들의 활용에 어려움이 있었다. 본고에서 소개한 안내서는 국내외 논의와 연구·정책 동향을 적극 반영하여 만든 실무 지침서로, 향후에는 안내서가 국내 기업들이 신뢰할 수 있는 인공지능을 구현하는 데에 도움이 되는 길잡이가 되길 바란다.

더 나아가서는, 안내서를 기반으로 기업 스스로가 규범 또는 가이드라인을 정립하여 신뢰성 제고를 위한 실천으로 이어지고, 기업의 각 전문가가 실천 결과 및 실천 과정에서 있었던 어려운 점들을 자발적으로 공유할 수 있어야 할 것이다. 이처럼 실효성 및 전문성을 갖춘 논의를 통해 사용자들이 올바르게 인공지능 기술을 활용할 수 있도록 유도하고, 기술에 대한 잘못된 인식이나 지나친 우려 등을 해소하여 인공지능 기술의 신뢰성이 사회 전반에 자리 잡는 데에 이바지하기를 기대한다. 

* 본 연구는 2021년도 정부(과학기술정보통신부)의 재원으로 수행된 연구임.(인공지능 신뢰성 기반조성)

주요 용어 풀이

- 인공지능(AI, Artificial Intelligence): 컴퓨터로 구현한 지능 또는 이와 관련한 전산학의 연구 분야
- 개인정보보호 규정(GDPR, General Data Protection Regulation): 유럽 의회에서 유럽 시민들의 개인정보 보호를 강화하기 위해 만든 통합 규정

참고문헌

- [1] UNESCO, Recommendation on the ethics of artificial intelligence
- [2] OECD, AI Principle
- [3] European Union(EU), General Data Protection Regulation
- [4] 과학기술정보통신부, 사람이 중심이 되는 인공지능을 위한 「신뢰할 수 있는 인공지능(AI) 실현 전략」
- [5] 과학기술정보통신부, 한국정보통신기술협회, 2022 신뢰할 수 있는 인공지능 개발 안내서
- [6] ISO/IEC TR 24028:2020 Overview of trustworthiness in artificial intelligence
- [7] High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI
- [8] 과학기술정보통신부, 인공지능(AI) 윤리기준
- [9] European Commission, The Assessment List on Trustworthy Artificial Intelligence
- [10] ISO/IEC TR 24029-1:2021 Assessment of the robustness of neural networks - Part 1: Overview
- [11] ISO/IEC 23894 Risk management
- [12] ISO/IEC TR 24027:2021 Bias in AI systems and AI aided decision making
- [13] Google, People + AI Guidebook
- [14] ETSI GR SAI 005 Securing Artificial Intelligence (SAI); Mitigation Strategy Report
- [15] OECD, Recommendation of the Council on Artificial Intelligence
- [16] WEF, Companion to the Model AI Governance Framework
- [17] 금융위원회, 금융분야 인공지능(AI) 가이드라인
- [18] 과학기술정보통신부, 정보통신정책연구원, 인공지능(AI) 윤리 자율점검표
- [19] 개인정보보호위원회, 인공지능(AI) 개인정보보호 자율점검표