https://doi.org/10.7840/kics.2023.48.12.1652

인공지능 신뢰성 확보를 위한 글로벌 정책 비교 및 국내 적용 방안 연구

김 진 민◆, 이 민 철*, 서 정 훈*, 신 용 태°

A Study on the Comparison of Global Policies for Trustworthy AI and the Application Method in Korea

Jinmin Kim◆, Mincheol Lee*, Junghun Seo*, Yongtae Shin°

요 약

인공지능의 급속한 발전으로 인해 사회 혁신이 가속화되며 인공지능의 긍정적 효과와 함께 사회적 역기능의 문제가 점점 우려되고 있다. 인공지능 알고리즘의 편향성은 사회적 불평등을 증폭시키며, 대규모 데이터를 이용한 딥러닝 기술은 개인정보 침해와 정보 유출을 우려하게 만들고 있다. 더욱이, 인공지능 기술이 사회 전반에 적용되며 예상되는 법적 책임에 대한 불명확성은 시스템의 안전성 문제, 결정 과정의 불투명성과 결부되어 기술에 대한 사회적 신뢰를 저하시키고 있다. 이러한 이슈에 대응하기 위해 전세계 국가들은 앞다퉈 인공지능 기술의 신뢰성을 확보하기 위한 정부 정책을 만들어 적용하고 있다. 따라서, 미국, 영국, 유럽연합 등 글로벌 선도국들의 인공지능 정책 특징 및 그 변화상을 비교 분석하는 것은 국내 실정에 적합한 정부 정책을 마련하는 데 많은 시사점을 제공해 줄 것이다. 이에, 본 논문에서는 주요 국가들이 시행하고 있는 인공지능 신뢰성을 확보하기 위해 시행 중인 다양한 정책들을 비교 분석하여 공통점과 차이점을 도출하였으며 이를 바탕으로 인공지능 기술의 긍정적인 영향력을 극대화하고 부정적인 영향은 최소화하기 위한 신뢰성 기반 인공지능 정책 수립방안에 대해 제언하였다.

Key Words : AI, Trustworthy AI, Policy, Technology policy

ABSTRACT

The rapid advancement of AI has accelerated societal innovation. Along with the positive impacts of AI, concerns about its societal dysfunctions are growing. The bias inherent in AI algorithms has the potential to exacerbate social inequalities. Moreover, deep learning technologies using massive data raise concerns about privacy. As AI technologies are applied across society, the ambiguity regarding legal responsibility, diminishes public trust in the technology. To address these issues, nations worldwide are proactively establishing unique governmental policies to ensure the reliability of AI technology. Consequently, comparing and analyzing the features and evolving aspects of AI policies and the EU will provide valuable insights for formulating suitable domestic policies. In this paper, we have analyzed various policies implemented by major countries to ensure the trustworthy AI, drawing out commonalities and differences. Based on this, we propose policy recommendations grounded in trustworthiness. to maximize the positive impact of AI technology and minimize its adverse effects.

◆ First Author : Graduate School of IT Policy and Management, Soongsil University, mins831@naver.com 정회원
° Corresponding Author : Professor, Soongsil University, shin@ssu.ac.kr, 정회원
* Graduate School of IT Policy and Management, Soongsil University, navymincheol@gmail.com; sepsjh@naver.com
논문번호 : 202308-044-0-SE, Received August 13, 2023; Revised September 15, 2023; Accepted September 18, 2023

Ⅰ. 서 론

인공지능 기술의 급속한 발전은 우리의 일상생활, 경제, 사회, 그리고 문화에 깊은 영향을 미치고 있다. 그러나 인공지능 기술이 산업과 사회에 빠르게 도입되면서 혁신을 창출하고 있지만 동시에 인공지능으로 인한 사회적 역기능도 증가시키고 있다. 특히, 인공지능 학습에 사용되는 데이터와 관련된 이슈는 사생활 침해와 정보 유출의 우려를 높이고, 법적 책임과 규제에 관해 명확하지 않은 기준 또한 법적 분쟁의 원인으로 대두되고 있다. 따라서, 인공지능 기술의 급속한 발전과 그에 따른 사회적 악영향을 통제하기 위해서는 정부가 신뢰성을 담보할 수 있는 인공지능 기술 촉진 정책을 만들어 나가야 한다.

인공지능 기술 발전에 따라 촉발되는 여러 문제들에 대응하기 위해, 주요 국가들은 이미 신뢰성 기반의 인공지능 기술 기준을 만들고 이를 국가 주요 정책으로 만들어 적극 추진해 나가고 있다. 또한, 주요 국가들간의 인공지능 경쟁이 점점 더 심화되고 있는 가운데, 인공지능의 신뢰성 확보는 국가 경쟁력과 안보를 높이는 데 결정적인 역할을 하게 될 것이다. 이에, 국내에서도 인공지능의 예상되는 역기능을 막기 위하여 미국· 유럽 등 주요 선진 기술 국가들의 정책 동향을 주시하고 있다. 그러나 아직 주요 관심이 기술적인 부분에 집중되어 있어 국가 정책적 관점에서 인공지능 기술을 바라보며 국내 실정에 적용할 수 있는 정책 개발에 관한 연구는 아직 초기 단계에 머물러 있다고 할 수 있다.

인공지능의 정책 제도화 방안 및 신뢰성 확보와 관련된 선행 연구를 먼저 살펴보면 다음과 같다. 오승환 외 11명은 주요 산업별 인공지능 활용에서의 문제점을 소개하고 중장기적 관점에서의 해결방안을 제안하였다[1]. 유재흥은 미국, EU, OECD를 중심으로 인공지능 정책 동향을 정리하였으며[2] 장창기 외 2명은 인공지능 윤리기준의 자율적 준수를 유도하는 방안을 제시한 바 있다[3], 또한, 인공지능 신뢰성 분야와 관련해서 김근형은 설명 가능한 인공지능의 개념과 관련된 기술 모델을 분석하였고[4] 김기연 외 2인은 ISO/IEC에서 진행 중인 인공지능 신뢰성 국제 표준화 활동을 소개하였으며[5], 양희태는 인공지능의 주요 안전성 이슈를 연구·발표하였다[6]. 인공지능의 법적인 모호성도 주요 연구이슈 중의 하나이다. 정소영은 형사사법 체계에서 인공지능 사용에 대한 유럽의회 결의안을[7], 김희정은 형사사법 관점에서의 인공지능 윤리 이슈를 분석하였고[8], 김한균은 인공지능 사회로 진화하며 인공지능 기반 의사 결정 방식의 적용으로 인한 법제화 추진 방향을 제안한 바 있다[9].

이런 가운데 주요 국가들이 추진 중인 인공지능 신뢰성 정책의 변화상을 확인하고 이에 대한 비교 분석을 통해 공통점과 차이점을 도출하는 것은 중요한 의미가 있다. 이러한 분석을 통해, 인공지능의 신뢰성 및 안전성 문제에 대한 근본적인 원인이 무엇인지 확인하고 이를 바탕으로 국내 정책을 효과적으로 수립할 수 있다. 이에 본 논문을 다음과 같이 구성하였다. 우선 제2장에서 선행 연구로 인공지능의 안전성 이슈와 관련하여 인공지능의 기술적 위험성과 기술 적용에 따른 사회적 위험성을 구분하여 설명한다. 제3장에서는 인공지능 기술을 선도하고 있는 미국, 영국, 유럽연합 등 주요국들의 인공지능 신뢰성 평가 프레임워크를 비교 분석하였다. 제4장에서는 이들 국가들의 주요 인공지능 정책들을 비교·분석, 공통점과 차이점을 도출한 뒤 이를 바탕으로 국내 인공지능 기술의 위험성을 완화하고 신뢰성을 확보하기 위한 법제 방향, 평가 인증체계 구축방안 등의 국내 정책 방안을 제언하였다.

Ⅱ. 관련 연구

2.1 인공지능 기술 기반의 위험성

인공지능 기술의 발달로 보편화되면서 많은 분야에서 인공지능이 활용되고 있다. 따라서 인공지능 기술이 적용된 정보시스템 자체를 대상으로 하는 악의적인 공격은 매우 치명적이다. 이에 인공지능 기술을 위협할 수 있는 여러 공격 기술에 대해 주목할 필요가 있다. 인공지능 기술 자체를 대상으로 공격하는 중독, 회피, 모델추출 공격기법 등의 위험성을 먼저 인식하고 있어야 한다. 이에 본 절에서는 학습 데이터를 조작하여 인공지능 모델의 정확도를 떨어트리는 중독. 인공지능을 속여 예측 결과를 왜곡시키는 회피 공격. 인공지능의 학습 결과를 무단으로 탈취하여 악용하는 모델추출 공격의 기본적인 동작 원리와 그로 인해 파급되는 부정적 영향을 살펴보고자 한다.

2.1.1 중독(Poisoning)

인공지능 중독 공격은 인공지능 모델의 학습 과정에 악성 데이터를 주입하여 인공지능 모델을 손상시키는 공격이다. 공격자는 인공지능 모델이 학습하는 데이터에 악성 데이터를 추가하거나, 기존의 데이터를 조작하여 인공지능 모델이 잘못된 결과를 도출하도록 유도한다. 데이터 중독은 잘못된 데이터를 제공하여 이를 학습한 인공지능 알고리즘이 잘못된 결과를 내도록 하는 것으로 최소한의 데이터로 최대한의 오동작을 일으키게

된다. 학습용 데이터에 악성 데이터를 삽입하는 방법을 사용하여 AI 학습모델의 딥러닝 과정에 관여하여 AI 시스템 자체를 손상시키는 공격이다[10]. 중독 공격이 발생하여 학습 데이터가 왜곡되거나 손상되면 시스템의 정확성과 신뢰성이 떨어지고 오류가 발생할 가능성이 커진다. 이로 인해 사용자들의 신뢰를 잃게 될 뿐 아니라, 인공지능의 신뢰성을 담보하는 '공정성'과 '안전성'에 주로 영향을 미치게 된다. 또한, 부정확한 의사 결정이 이루어질 수 있다. 중독 공격에 취약한 인공지능 시스템은 예측이나 추천 결과가 부정확해질 가능성이 있으며 결정 과정에서 인공지능의 '투명성'에도 영향을 미치게 된다. 다음으로 개인정보 침해가 발생할 수 있는데 중독 공격에 취약한 인공지능 시스템을 통해 민감한 개인정보가 유출될 위험이 생기며 이는 사용자의 프라이버시 침해로 법적 문제를 야기한다.

2.1.2 회피(Evasion)

인공지능 기술을 대상으로 하는 회피 공격은 인공지능 시스템의 입력 데이터에 악의적인 변조를 가해 정상적인 입력 데이터로 인식하도록 유도하여, 시스템의 성능을 손상시키는 공격이다[10]. 예를 들어, 스팸 필터링을 위한 인공지능 시스템을 회피하는 공격은, 스팸 메일에 추가한 단어를 불완전하게 표기하거나, 스팸 메일의 텍스트를 이미지 형태로 변환하는 등의 방법을 사용한다. 회피 공격은 인공지능 시스템의 모델 구조나 입력 데이터 등에 대한 깊은 이해를 기반으로 수행되는데 공격자는 시스템의 모델 구조와 학습 데이터에 대한 정보를 수집하여, 오류를 유발하는 새로운 입력 데이터를 생성한다. 이러한 입력 데이터는 시스템이 일반적으로 사용하는 데이터와는 다른 특성을 가지며, 시스템의 성능을 손상시킨다. 따라서, 회피 공격은 인공지능 시스템의 '안전성'을 저해시키는 가장 큰 요인 중 하나이다. 회피 공격을 통해 시스템이 잘못된 결정을 내리거나 예기치 않은 동작을 수행할 가능성이 커지게 되는데, 이는 인공지능의 '투명성'과 '설명 가능성'에 큰 악영향을 주게 된다. 또한, 회피 공격을 통해 공격자는 인공지능 기반의 보안 시스템을 속여 악성코드를 숨기거나, 시스템을 침입하는 데 성공할 수 있어 '안전성'에도 큰 위협이 된다. 이러한 회피 공격으로 인한 인공지능 시스템의 오작동은 사람들의 인공지능 기술에 대한 신뢰를 저하시키는 주요인이 된다.

2.1.3 모델 추출(Model extraction)

인공지능 모델추출 공격은 인공지능 모델의 파라미터를 추출하여 악의적인 목적으로 사용하는 공격이다. 공격자는 인공지능 모델이 학습한 데이터에 접근하여 모델의 파라미터를 추출하거나, 인공지능 모델의 동작을 분석하여 모델의 파라미터를 추측할 수 있다[11]. 일반적으로 모델추출 공격은 머신러닝 모델을 대상으로 한다. 보다 구체적으로 살펴보면 공격자는 인공지능 모델에 입력을 제공하여 모델의 출력 결과를 수집하고 이 출력 결과는 모델의 내부 동작 및 구조를 추정하는 데 사용될 수 있다. 또한, 공격자는 대상 모델에 대한 다양한 입력을 제공하여 모델의 응답을 분석하고 모델이 사용하는 알고리즘 및 파라미터에 대한 정보를 수집하고 이렇게 수집한 정보를 기반으로 공격자는 대상 모델의 복제본을 생성하거나 모델을 해킹하여 공격을 수행하는 방식이 가능하다. 따라서, 모델 추출 공격은 인공지능 모델의 보안을 취약하게 만든다. 예를 들어, 의료 분야에서 인공지능 모델이 모델 추출 공격을 당하면 환자의 의료 기록이 유출될 수 있고, 이에 따라 개인정보가 유출될 가능성이 생기게 되는데 이는 모델의 '개인정보보호' 및 '보안성'에 문제를 발생시킨다.

2.2 인공지능 기반 사회가 초래하는 위험성

인공지능 시스템이 사회 각 영역에 적용됨으로써 생기게 되는 사회적 영향 및 그 위험성 또한 고려해야 한다. 인공지능 기술을 기반으로 형성된 이른바 인공지능 플랫폼 생태계는 관련 기술을 연구하는 연구자, 관련 시스템을 개발 생산하는 기업, 이를 이용하는 일반 사용자, 그리고 이들이 구성원으로 참여하게 되는 사회, 국가에 이르기까지 깊은 영향을 주게 된다. 본 절에서는 인공지능 기술의 대표적인 사회적 역기능으로 알려진 편향성, 불투명성, 공정성, 보안성 이슈 등에 대해 발생원인 및 파급영향에 대해 살펴보고자 한다.

2.2.1 편향성

먼저, 인공지능 알고리즘의 편향성 문제는 학습 데이터의 불균형과 개발자의 편견에서 기인하며, 이에 따라 인종, 성별, 나이 등의 차별이 발생할 수 있다. 대표적 예를 들어 보자면 2000년 인공지능 기반의 챗봇 서비스 '이루다' 시스템이 출시 한 달 만에 서비스를 중단한[12] 사례가 있다. 이는 개인정보 침해, 성별에 관한 혐오, 외설적 목적 사용 등의 부작용에 따른 것이었다. 인공지능 채팅봇의 작동 원리가 대량의 학습 데이터를 활용한 딥러닝과 자연어 처리 기술을 기반으로 작동하는데, 대량의 데이터를 이용하여 모델을 학습시키고, 학습된 모델을 활용하여 새로운 데이터를 예측하는 방식이다. 자연어 처리 기술은 기본적으로 인간의 언어를 이해하고 처리할 수 있는 기술로, 인공지능 채팅봇에서는 사용자

의 입력을 이해하고 적절한 대답을 생성하기 위해 이용된다. 그러나 인공지능 채팅봇은 학습 데이터에 의존하기 때문에, 학습 데이터에 포함된 편향성이 채팅봇의 대화에 영향을 미칠 수 있다. 예를 들어, 특정 인종이나 성별을 대상으로 하는 편견이 학습 데이터에 반영되어 있으면, 인공지능 채팅봇이 이러한 편견을 반영한 대답을 생성할 가능성이 커진다. 또한, 인공지능 채팅봇이 대화를 생성할 때, 이전 대화에서의 내용을 참조하는 경우가 많은데 이때, 이전 대화에서의 편향성이 다음 대화에서도 계속해서 반영될 수 있다. 잘못된 학습 때문에 유발되는 편향성 문제는 사회적 분열, 신뢰도 저하, 법적 책임 및 기술 혁신 저해와 같은 부정적 영향을 미친다. 이 문제를 해결하기 위해서는 다양한 배경과 경험을 가진 개발자들이 협력하여 시스템을 설계하고, 균형 잡힌 학습 데이터를 사용하는 것이 중요하다. 인공지능 편향성 문제를 해결하기 위해, 데이터 수집, 학습 알고리즘 개선, 대화 기록 관리 등 다양한 기술과 정책적인 요소들이 필요하다.

2.2.2 불투명성(설명 가능성 부재)

인공지능은 학습 과정에서 많은 데이터를 학습하고 이를 기반으로 판단한다. 이는 인공지능이 어떻게 작동하는지, 어떤 데이터를 기반으로 의사결정을 내리는지 인간이 이해하기 어렵다는 것을 의미한다. 인공지능은 대규모 데이터를 학습하여 의사 결정을 내리기 때문에, 인간이 이해하기 어려운 복잡한 알고리즘을 사용하며 학습 과정에서 인간이 거의 개입하지 않기 때문에 어떻게 의사 결정이 내려졌는지 그 근거를 명확히 설명하기 어렵다. 이러한 결정 과정의 불투명성 문제는 사람들이 그 과정을 이해하거나 추적하기 어렵다는 점에서 '인공지능의 블랙박스' 문제로도 알려져 있다. 이러한 문제로 인해 인공지능의 편향성, 책임소재 모호성 등의 문제가 발생한다. 설명 가능한 인공지능은 결정 과정의 투명성을 확보하여 사용자와 사회의 신뢰를 얻고, 법적, 윤리적 문제를 해결하는 데 기여할 수 있다. 특히, 인공지능 기반 추천 시스템은 주로 사용자의 이전 행동 기록과 같은 데이터를 바탕으로 작동하는데 이러한 데이터를 바탕으로 추천 시스템은 사용자의 취향과 관심사를 파악하고, 그에 따라 해당 사용자가 선호할 만한 제품, 서비스, 콘텐츠 등을 추천한다. 이때 추천 시스템은 다양한 알고리즘 기술을 사용한다. 주요한 알고리즘으로는 Content-based Filtering, Collaborative Filtering, Hybrid Filtering 등이 있다[13]. 그러나, 사용자에게 왜 해당 콘텐츠가 추천되었는지에 대한 이유를 설명하지 못하기에 일관성이 없거나, 추천 결과의 생성 과정을 설명할 수 없다.

2.2.3 불공정성

최근 인공지능 기술이 초래하는 공정성 문제에 대한 사회적 우려 또한 중요한 고려 요소이다. 인공지능의 불공정성은 인공지능 기술이 사회 전반에 널리 사용됨에 따라 더욱 심각한 문제로 부각되고 있는데 이는 개인의 권리 침해, 사회적 불평등 및 갈등 유발의 가능성이 있기 때문이다. 불공정성 문제는 주로 데이터 편향, 설계자의 편견, 비표현적 특징 등이 원인으로 발생한다. 훈련 데이터가 특정 집단에 편향되거나, 설계자가 자신의 가치관이나 편견을 알고리즘에 무의식적으로 반영할 때, 인공지능의 판단 결과가 공정하지 않게 될 수 있다. 예를 들어, 미 NIST 연구 결과에서 전세계 189개 안면 인식 알고리즘을 조사한 결과, 대부분이 백인 데이터에 최적화되어 있어, 다른 인종이나 민족에 대한 인식률이 상대적으로 낮다는 결과가 나왔다[14]. 이는 인공지능 기술의 편향성과 공정성 문제를 더욱 부각시키는 결과로 나타났다. 공정한 인공지능 시스템은 개인의 특성이나 그룹 구성과 관계없이 동일한 기준을 적용하고, 비슷한 상황에 놓인 이들에게는 동일한 결과를 제공해야 한다. 인공지능의 편향성 문제는 데이터와 알고리즘의 균형을 맞추는 것이 목표이고, 공정성은 그 결과로부터 나오는 판단과 처리가 공평한지를 확인하고 조정하는 것이 목표라 할 수 있다. 이 두 문제는 서로 연관되어 있지만, 다루는 측면과 접근 방식에 차이가 있다고 할 수 있다.

2.2.4 보안성·안전성(통제 가능성)

인공지능 기술의 급격한 발전은 많은 긍정적인 영향을 미치지만, 동시에 사이버공격 및 해킹 문제에 대한 우려 또한 높아지고 있다. 인공지능 기술이 악용되어 발생할 수 있는 사이버 보안 위협으로는 악성코드 자동생성, 사회공학적 해킹방식의 고도화, 보안프로그램 탐지 회피, 악성코드 확산의 손쉬움 등이 있다. 인공지능은 악성코드를 빠르고 효율적으로 생성하며, 전통적인 보안 솔루션을 우회할 능력을 갖추고 있다. 또한, 인공지능을 활용한 사회공학적 공격은 피해자의 정보를 분석해 맞춤형 공격 메시지를 생성해 전송함으로써 더욱 정교하게 작용한다. 이 외에도 인공지능은 보안 시스템의 행동을 모방하여 악성코드를 변조하거나 탐지를 피할 수 있다. 특히 최근에는 '챗GPT'와 같은 생성형 인공지능에 악의적인 목적을 가지고 답변을 유도하는 기술이 주목받고 있는데 이를 전문으로 하는 '탈옥(Jail breaking) 프롬프트' 같은 사이트까지 등장했다[15]. 또

한, 인공지능 기술의 발전은 대규모 자료수집과 처리에 의존하며, 이에 따라 개인정보 침해와 데이터 유출 위험이 발생한다. 무분별한 자료수집과 처리는 개인의 프라이버시를 침해할 우려가 있는데 대표적인 예로 카카오톡 이용자 60만명의 대화를 무단 사용한 '이루다' 사례[16]를 들 수 있다.

Ⅲ. 인공지능 기술 신뢰성 평가

3.1 인공지능 신뢰성 평가 기준

인공지능 기술이 사회에 전면적으로 도입되면서 세계 각국에서는 다양한 인공지능 정책 및 윤리기준을 만들어 발표하고 있다. 인공지능 기술의 신뢰성을 평가하기 위해서는 우선 누구나 이해할 수 있는 신뢰성 평가 기준을 정의하고 이를 투명하게 공개하는 것이 중요하다. 이렇게 표준화된 인공지능 신뢰성 평가 기준 또는 체계(프레임워크)는 정부 정책에 대해 신뢰성을 부여해 주며 인공지능 기반 시스템의 개발·배포·사용을 위한 기준 역할도 하게 된다. 따라서 다양한 정부에서 각자의 정책 방향에 맞춰 인공지능 평가 프레임워크를 개발하고 있다. 이를 통해 인공지능 시스템이 투명하고 공정하며 신뢰할 수 있게 설계·사용되고 있는지 확인할 수 있으며 기술의 안전성·신뢰성을 보장해 줌으로써 인공지능에 대한 사회적 우려를 감소시킬 수 있다. 따라서 본 장에서는 유럽연합, 미국, 영국이 개발한 대표적인 인공지능 평가체계 5종에 대해 분석하고 비교해 보았다.

3.2 유럽연합(EU)의 ALTAI[17]

유럽연합은 2020년 신뢰할 수 있는 인공지능 시스템을 구축하기 위한 자가 평가 방식의 'Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment[17]'를 발표하였다. 이 프레임워크는 인공지능 시스템의 개발, 배포, 사용의 모든 단계에서 인공지능 시스템의 안전성, 공정성, 윤리성, 효율성, 영향력 등을 체계적으로 평가하기 위한 목적으로 개발되었다.

ALTAI는 유럽연합의 윤리 지침을 바탕으로 개발되었으며, 다음<표 1>과 같은 7개의 주요 평가 요소들로 구성되었다. 무엇보다도 인간 중시(Human Agency and Oversight)를 첫 번째 요구사항으로 내세우며 인공지능 시스템은 인간에 의해서만 감독 되어야 하며 책임, 의사결정 권한 등이 궁극적으로 인간이 보유하도록 설계되어야 한다는 점을 최우선적으로 강조하고 있다. 이를 기반으로 인공지능 시스템 개발·운영 각 과정에서 안전성 검사, 시스템 장애 대응 등을 통하여 기술 안정성과 견고성(Technical Robustness and Safety)을 확보하고,

표 1. ALTAI의 주요 평가 요구사항 및 평가 요소
Table 1. Major Requirement & Evaluation Criteria of ALTAI

Requirement	Evaluation criteria
1 Human Agency and Oversight	1. Human Agency and Autonomy 2. Human Oversight
2 Technical Robustness and Safety	1. Resilience to Attack and Security 2. General Safety 3. Accuracy 4. Fall-back plans and Reproductibility, Reliability,
3 Privacy and Data Governance	1. Privacy 2. Data Governance
4 Transparency	1. Traceability 2. Explainability 3. Communication
5 Diversity, Non-discrimination and Fairness	1. Avoidance of Unfair Bias 2. Accessibility and Universal Design 3. Stakeholder Participation
6 Societal and Environmental Well-being	1. Environmental Well-being 2. Impact on Work and Skills 3. Impact on Society at large or Democracy
7 Accountability	1. Auditability 2. Risk Management

인공지능의 작동 방식, 데이터 활용 등에 대한 충분한 설명을 사용자에게 제공하고 결과에 대한 차별이 발생하지 않도록 다양한 인적 특성(성별, 인종, 나이, 지역 등)을 반영하는, 편향성 감시 기능을 갖춤으로써 투명성(Transparency)과 공정성(Fairness)을 갖춰야 한다고 설명하고 있다 이와 더불어, 사회 이익과 가치를 존중하기 위한 철학과 함께 법적인 규제 준수, 사용자 개인정보를 보호하고 적절한 데이터 관리가 이루어져야 한다. 이어서 시스템을 개발·운영하는 모든 주체가 인공지능 설계 및 사용에 책임을 져야 한다. 인공지능을 개발하거나 활용하려는 유럽연합의 기관들은 해당 7개 기준을 충족하는지를 자체 검사하도록 권고되고 있다.

3.3 영국의 'Data Ethics Framework'

'Data Ethics Framework[18]'는 영국 정부가 공공 부문에서 데이터 사용과 인공지능 시스템을 이용할 때의 기준으로 개발되었으며 2018년 최초 발표되었고 이어 2020년 개정판이 공개되었다. 이 프레임워크는 영국의 데이터 윤리 기본 지침서 역할을 하며 해당 프레임워크는 <그림 1>과 같이 투명성, 책임성, 공정성 3가지를 주요 원칙을 가지고 각각의 달성도를 0에서 5까지 점수

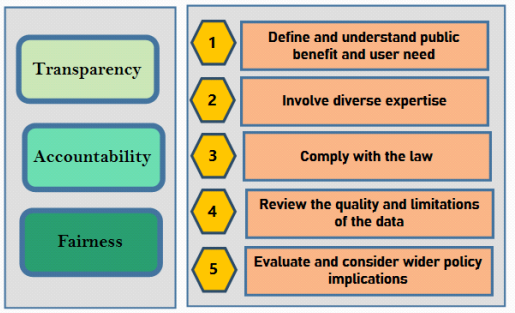


그림 1. '데이터윤리 프레임워크'의 주요 원리 및 단계
Fig. 1. Principles and stage of 'Data ethics Framework'

로 매기도록 구성하였다.

주요 원칙에서 우선 투명성(Transparency) 원칙에 따라 인공지능 프로젝트에 대한 정보를 이해하기 쉽게 접근할 수 있도록 개발사 프로세스 및 데이터를 대중에 공개하고 책임성(Accountability) 원칙에 따라 효과적인 시스템 감독 절차를 가져야 하는 데 특히 공공의 목적으로 개발될 시에는 정부에 의해 효과적인 감독과 통제가 수행되어야 한다. 마지막으로 공정성(Fairness)을 위해 개인과 사회 집단에서 의도하지 않게 차별을 만들 가능성을 제거하는 것인데 인공지능 학습모델 결과에 영향을 미칠 수 있는 편견을 완화해야 하고 그 결과가 개인 인권과 민주적 가치에도 일치해야 함을 강조하고 있다. 영국의 'Data Ethics Framework'는 위에서 설명한 3가지 원칙을 기반으로 인공지능 프로젝트 개발에 사용된 데이터가 얼마나 정확하고 대표적으로 사용되었는지 확인하고 사용된 데이터가 정당한 책임을 지고 활용되고 있는지 지속적으로 평가하기 위한 평가체계를 정의하고 있다. 인공지능 기반의 데이터 시스템을 개발·활용하려는 기관들은 위의 평가 단계 어느 하나에서라도 3점 이하의 점수를 받을 경우, 프로젝트 개발 단계에서 이를 개선하기 위한 추가적인 검토가 이루어져야 한다. 특히, 점수가 낮게 나온 이유를 반드시 설명하고 윤리적 표준을 개선하기 위한 구체적인 조치를 수행하여야 한다.

3.4 영국의 'AI Auditing Framework'

영국의 데이터 보호와 개인정보 보호를 감독하기 위하여 만들어진 정부 기관인 ICO(Information Commissioner's Office)는 'AI Auditing Framework[19]'를 2020년 발표하였다. 영국 정부는 인공지능 시스템의 개발과 사용에서의 데이터 보호와 관련된 위험을 이 프레임워크를 통해 식별, 평가, 관리하도록 권고하고 있다.

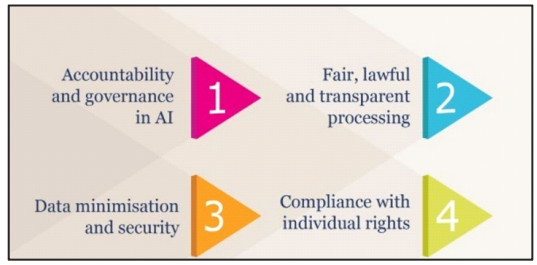


그림 2. 영국 'AI 감사 프레임워크'의 절차
Fig. 2. Procedure of 'AI auditing Framework'

영국 ICO는 'AI Auditing Framework'를 크게 6가지 분야로 분류하여 시스템을 검사하도록 권고하고 있는데 합법·공정성 및 투명성, 목적 제한(Purpose limitation), 데이터 최소화(Data minimisation), 정확성, 데이터 저장 기간 제한, 보안 책임 등이다. 'AI Auditing Framework'에서 눈여겨볼 대목은 <그림 2>와 같이 4가지 절차를 걸쳐 데이터의 사용에 대해 엄격히 검사하고 있다는 점이다. 예를 들어 목적 제한 분야에서는 인공지능 개발에 사용된 데이터는 원래 수집된 목적과 일치 또는 호환되는 범위 내에서만 처리되고 있음을 검사하고 데이터 최소화 및 데이터 저장 기간 제한 항목을 통해 시스템 개발 목적을 달성하기 위한 최소한의 필요 데이터만 사용하는 한편 인공지능 개발 및 운영에 사용된 데이터는 개발 목적 달성을 위해 필요한 기간만 저장하도록 권고하고 있다.

3.5 미국의 'AI: An Accountability Framework for Federal Agencies and Other Entities[20]'

2021년 미국 회계 감사원 GAO(Government Accountability Office)는 인공지능 시스템을 책임감 있게 사용하기 위한 실용 가이드라인으로써 'Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities[20]'을 발표했다.

이 정책 문서에서 연방 기관과 기타 공공조직들이 인공 지능 시스템을 책임감 있게 사용하기 위한 기준 평가 모델을 공개하였는데 이 프레임워크는 <그림 3>과 같이 크게 데이터, 거버넌스, 성능, 모니터링 등 네 가지 구성 요소로 이루어져 있다. 우선 인공지능 기술을 개발, 운영, 유지 보수를 하기 위해서는 그에 맞는 거버넌스 체계가 마련되어야 하며 시스템은 기본적으로 데이터를 기반으로 작동하기 때문에 데이터의 품질, 편향성, 개인정보보호 등 다양한 측면에서 평가되어야 한다. 또한, 인공지능 시스템이 지속적인 성과를 발휘할 수 있도록 시스템 작동 상태나 데이터의 변화 등을 지속적

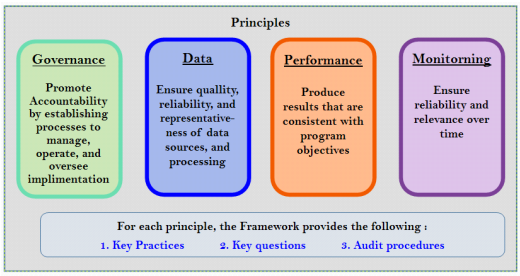


그림 3. 'GAO 인공지능 프레임워크'의 4가지 주요 요소
Fig. 3. Four Components of 'GAO AI Framework'

으로 모니터링하고 평가함과 동시에 시스템의 성과를 지속적으로 평가하고, 개선 방안을 모색해야 한다. 미국 GAO가 발표한 해당 문서는 인공지능 시스템의 내부 투명성(기관 내에서의 이해)과 외부 투명성(공공의 이해)의 중요성을 강조하며 동시에 인공지능 시스템을 설계하고 운영하는 데 있어서의 윤리적, 법적, 정책적 요구사항들을 별도로 정의하고 있다.

3.6 미국의 'AIRMF'

미국 국립표준기술연구소(NIST)가 2023년 발표한 'AI Risk Management Framework(AIRMF)'는 미국의 인공지능 시스템의 효과적인 관리를 위해 인공지능의 위험을 식별하고 관리하는 방법론을 제시하였다[21].

NIST는 인공지능 시스템의 설계, 개발, 배포, 사용에 있어 신뢰성을 평가할 수 있는 표준 방법론으로써 해당 프레임워크 준수를 권고하였는데 <그림 4>와 같이 거버넌스(Govern), 관리(Manage), 측정(Measure), 지도(Map) 등 네 개의 위험 관리 목표를 제시하였다. 첫 번째 목표인 거버넌스(Govern)는 인공지능 시스템을 설계, 개발, 또는 이용하는 조직은 조직 내에서 인공지능 위험을 관리하기 위한 프로세스 및 문화를 조성해야 하는데 이를 위해서 조직 내 AI 위험을 매핑(Map), 측정(Measure), 관리(Manage)하기 위한 정책이 투명하게 시행되어야 하고 다양성, 공정성을 위해 인종, 성별, 경력, 전문성, 배경 등 다양한 측면이 조직내에서 고려되어야 한다. 또한, 제3자 소프트웨어(SW), 데이터 및 기타 다른 공급망과 관련하여 발생하는 위험을 적절히 처리되어야 한다. 두 번째 지도(Map) 항목에서는 인공지능 시스템과 관련된 위험을 프레임화하기 위한 기본적인 관계성이 적절히 식별되어 있는지 확인하는 절차로 Map 기능의 결과는 뒤에 설명하는 측정(Measure)과 관리(Manage) 기능의 기초 역할을 하게 된다. 이 목표를 달성하기 위해서는 조직의 미션 및 인공지능 개발 목표 설정, 시스템을 구현하는 데 사용되는 방법(예: 분류기, 생성 모델), 인공지능 시스템 도입의 예상 이익 및 비용 분석 등이 있다. 세 번째 측정(Measure) 항목에서는 인공지능 위험과 관련된 영향을 분석, 평가 및 모니터링하는 요소로 적절한 측정 지표를 만들어 식별·적용함으로써 인공지능 시스템의 신뢰성 특성을 평가하고 인공지능의 예상되는 위험을 추적하는 한편, 측정 항목의 효과성에 대한 피드백도 수집해야 한다. 마지막으로 관리(Manage) 항목에서는 지도(Map) 항목에서 확인한 시스템의 기초 정보 및 측정(Measure)에서의 분석·평가된 정보를 활용하여 인공지능 위험을 식별하고 인공지능 기술의 역효과 발생 가능성을 감소시키도록 하고 있다. 미국 인공지능 시스템의 위험 관리를 위한 표준적인 프레임워크로 개발된 AIRMF가 공개됨으로써, 미국에서의 인공지능 시스템을 평가 관리하기 위한 표준적인 방법론으로 동 프레임워크가 개발, 보급되고 있는데 이러한 미국의 정책 접근법은 인공지능의 안전성과 신뢰성 확보를 통한 인공지능 기술의 사회적 수용을 증진하는 데 크게 도움이 될 것으로 예상된다.

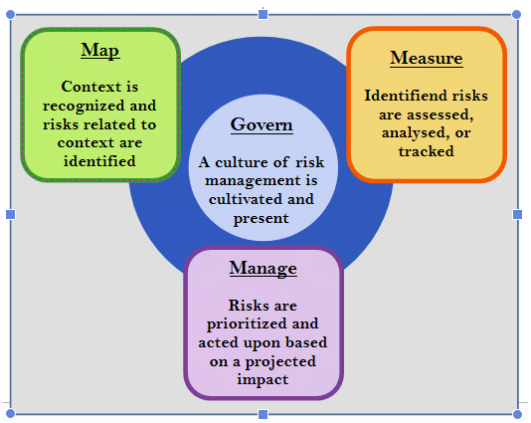


그림 4. 'AI RMF'의 4가지 코어 요소
Fig. 4. Four Core elements of 'AI RMF'

3.7 소 결

인공지능 기술이 급속도로 발전함에 따라 인공지능 시스템의 잠재적 위험에 대한 우려가 증가하고 있다. 인공지능 시스템이 잘못 사용될 경우 개인 프라이버시 침해, 차별, 편향 등 다양한 문제를 발생시킬 수 있기 때문이다. 이에, 인공지능 시스템의 신뢰성을 확보하기 위해 체계적이며 표준적인 평가 프레임워크가 필요하다. 이러한 신뢰성 평가 기준은 시스템의 잠재적 위험을 식별하고 관리하기 위한 기준점이 되며 또한, 시스템이 안전하게 개발, 사용, 운영되도록 보장함으로써 사람들이 인공지능 관련 기술을 더욱 신뢰하게 됨으로써 기술의 사회적 수용성을 증대시키고 기술의 적용과 확산에

중요한 역할을 하게 된다. 위와 같은 이유들로 인해, 유럽연합, 영국, 미국 등의 기술 선도국들은 인공지능의 신뢰성을 확보하고 책임성을 명확히 하기 위한 평가 프레임워크를 앞다퉈 개발·보급하고 있다. 위에서 살펴본 바와 같이 유럽연합, 영국, 미국 등 기술 선도국들이 발표한 인공지능 신뢰성 기반 평가 프레임워크는 각각의 특징을 지니고 있다. 유럽연합의 'Assessment List for Trustworthy Artificial Intelligence for self-assessment[17]' 프레임워크는 '신뢰할 수 있는 AI'에 중점을 두고 AI 시스템이 유럽연합의 윤리 가이드라인을 따르고 있는지를 기준으로 신뢰성을 평가 확인하고 있었으며 영국의 'Data Ethics Framework[18]'는 데이터를 처리하고 활용하는 방법론에 주로 중점을 두고 인공지능을 포함한 모든 데이터 기반 시스템이 책임감 있게 사용되어야 함을 강조하고 있다. 반면, 미국 GAO의 'Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities[20]'는 미국 연방 기관들에 적용되기 위한 기술 가이드라인으로 인공지능 시스템의 설계, 개발, 배포, 사용 과정에서의 책임성을 강조하고 있다. 또한, 미국 NIST의 'AI Risk Management Framework(AIRMF)[21]'는 인공지능의 리스크 관리에 중점을 두고 공공 부문과 민간 부문 모두에서 사용될 수 있도록 설계되었다는 특징을 가지고 있었다.

Ⅳ. 글로벌 AI 정책 비교 및 국내 적용 방안

앞 장에서 살펴본 바와 같이 공개된 각국의 인공지능 정책을 비교 분석하는 것은 국내에 적합한 제도나 지침을 마련하는 데 좋은 시작점이 될 것이다. 특히, 각 국가와 기관들이 어떻게 자신들만의 사회문화적, 경제적, 정치적 맥락 속에서 인공지능 기술을 해석하고 적용하고 있는지를 이해하는 것은, 국내 환경에 적합한 정책을 수립하는 데 있어 의미 있는 통찰력을 제공해 줄 것이다. 본 장에서는 인공지능 신뢰성 확보를 위해 추진된 주요국들의 공개 정책을 비교 분석하여 주요 공통점과 차이점을 확인하였다. 또한 이를 바탕으로 인공지능 신뢰성 향상을 위한 국내 정책 수립 방향에 대해 5가지 분야로 나눠 제안해 보았다.

4.1 국내외 주요 정책 비교 및 평가

미국, 유럽연합, 영국, 호주 등이 공개한 인공지능 정책을 비교해 보면 그 특징을 크게 두 갈래로 나누어 볼 수 있다. 구글, 애플, 마이크로소프트, OpenAI 등 다수의 글로벌테크 기업을 보유한 미국은 시장주도, 기업 친화적 접근 방식을 따르고 있으며 유럽연합, 영국, 호주 등 유럽 문화 국가들은 인간 가치 존중, 개인정보 보호를 우선시하는 규제 접근 방식을 택하고 있다는 점이다.

먼저, 미국의 인공지능(AI) 정책 특징을 살펴보면 '시장 중심과 기술 친화적'이라 설명할 수 있다. 이러한 접근 방식은 인공지능 기술의 적용 및 활용을 적극 장려하여 기술 혁신이 자연스럽게 촉진되도록 장려함으로써 자연스럽게 미국의 경제적 이익과 미국의 기술 패권, 기술 안보를 확보하려는 노력의 일환이라고 분석해 볼 수 있다. 그 예로 미국 상공회의소는 인공지능 기술의 상용화가 가시화되자 2019년 9월 50개가 넘는 상공회의소 회원사들과 협력하여 인공지능 기술의 책임 있는 사용과 이를 규제하기 위한 'AI 윤리 원칙(AI Principles)[37]'을 발표하였다. 또한, 2019년 트럼프 대통령이 발표한 'AI 이니셔티브(American AI Initiative)[22]'의 주요 내용을 살펴보면 미국의 AI 연구와 개발 촉진을 주요 목표로 선정하고 인공지능 기술에 대한 연방 투자 증가, 인공지능 기술을 활용하고 개발하는 데 필요한 규제 장벽 제거에 그 목표를 두고 있었다. 이를 뒷받침하기 위해 미 의회는 2021년 국가 'AI 이니셔티브 법안 [23]'을 채택하여 국가 차원의 법적 틀을 제공하였다. 미국의 기술 정책을 이해하는데, 있어 미국의 신기술을 선도하고 표준을 이끌어 가는 미국 국립표준기술원(NIST)의 정책 방향을 주의 깊게 살펴볼 필요가 있다. 2019년 12월 미국 NIST는 '얼굴인식 알고리즘에 관한 의미 있는 연구 결과 보고서를 발표하였는데 이 연구에서 전 세계적으로 사용되는 189개의 안면 인식 알고리즘에서 대부분 백인 데이터에 최적화되어 있어 타 인종에 대한 인식률이 급격히 떨어진다고 발표하며 인공지능 기술의 공정성 문제를 공개 제기하였다[14]. 이어 NIST는 2021년 10월 'Technical AI Standards', 즉 인공지능 기술의 개발과 사용을 위한 기술 표준을 발표하였다. 해당 표준은 AI 시스템 개발자, 사용자, 규제 기관 등 다양한 이해관계자를 대상으로 하는 표준 지침 역할을 하고 있다[24]. NIST가 발표한 표준은 투명성, 공정성, 보안성을 고려하면서도, 기업들이 AI를 활용하는 데 있어 과도한 규제로부터 자유롭게 하기 위한 목적 또한 포함하고 있다. 이러한 접근 방식은 인공지능 기술이 근본적으로 미국의 시장 경제적 가치를 증가시키고 사회적 문제를 해결하기 위한 주요 방편으로 인식하고 있기 때문이다. 이처럼 미국의 인공지능 정책은 대통령 행정명령과 그와 연계된 법률, 그리고 국가 기관 및 NIST의 기술 표준 공개를 기반으로 발전해 가고 있으며 이를 통해 미국은 인공지능 분야에서 세계적인 리

더 역할을 유지하고자 하였다.

다음으로, 유럽연합(EU), 영국, 호주 등 유럽 문화권의 인공지능 정책의 특징이다. 유럽 문화권에서는 주로 인간의 가치와 권리 존중, 개인정보 보호를 중심으로 한 접근 방식을 따르고 있다. 이는 "사람 중심의 AI"를 목표로 하는 유럽의 기본원칙에서도 알 수 있다. 특히 전세계 최초로 개인정보 보호 규정(GDPR)을 만든 유럽연합에서는 인공지능과 데이터 활용에 있어서 데이터 보호와 개인의 프라이버시를 강조하는 문화가 형성되었다. 2018년 EU에서 발표한 '인공지능을 위한 유럽 전략(AI for Europe)[25]'을 살펴보면 유럽 내에서 인공지능을 이용한 혁신을 촉진하면서도 인간의 가치와 윤리적 원칙을 존중한다는 가장 중요한 원칙으로 내세웠다는 것을 확인할 수 있으며 2019년 EU가 다시 '신뢰할 수 있는 인공지능을 위한 윤리 가이드라인(Ethics guidelines for trustworthy AI)[26]'을 발표하면서 인공지능이 인간의 권리와 자유를 존중하며, 유리하고 투명하게 작동해야 한다는 원칙을 다시 한번 강조한 바 있다. 이 지침은 향후 인공지능 시스템의 개발과 사용에 대한 유럽의 윤리기준 역할을 하게 된다.

호주 정부와 영국 정부도 2019년 발표한 'AI Ethics Framework and Principle'과[27] 2020년 발표한 'National Data Strategy'[28]에서 인공지능 기술의 사용을 책임감 있고, 윤리적으로 이루어져야 함을 명시하며 인공지능이 개인의 권리를 존중하고 사회적 가치를 높이는 데 기여한다는 점을 똑같이 강조한 바 있다. 마지막으로 2021년 유럽연합은 인공지능에 대한 종합적인 규제방안을 정리한 '인공지능 법안(AI Act)[29]'을 공개했는데 이는 매우 의미가 크다. 유럽의 개인정보보호 규정(GDPR)이 초창기 규제 거버넌스를 구체화함으로

표 2. 주요 국가들 및 국제기구별 인공지능 윤리 원칙 비교
Table 2. Comparison of AI Ethics Principles in Major Countries & International Organization

U.S. Artificial Intelligence Principles[37] (2019)	EU Ethics guidelines for trustworthy AI[26] (2019)	OECD AI Principles[31] (2022)	Korea's National AI Ethics Standards[33] (2020)
AI-Ready Workforce	Accountability	Accountable	Accountability
Cross-Border Data Flows	Fairness	Human rights, fair	Data Management
International Standards	Human oversight	Robust, secure and safe	Privacy Protection
Open and Accessible Data	Privacy and data governance	Transparency	Prohibition of Infringement
Partnership	Robustness and safety	Well-being.	publicity
Private and Public Investment	Transparency, traceability		Respect for diversity
Promote Innovation	Well-being		Safety
Risk-Based Approaches			Solidarity
Robust and Flexible Privacy			Human rights
Rules and Regulations			Transparency
U.S. Trustworthy Artificial Intelligence[38] (2020)	U.K. understanding AI ethics and safety[39] (2019)	IEEE AI Principles[30] (2022)	Australia's AI Ethics Principles[27] (2019)
Accountable	Accountability	Human right	Accountability:
Accurate, reliable	Fairness	Well-bing	Contestability
Lawful and Nation's values.	Sustainability	Accountability	Fairness
Performance-driven	Transparency	Transparency	well-being
Regularly monitored		Dealing with misuse	Human-centered values
Responsible and traceable			Privacy and security
Safe, secure, and resilient			Reliability and safety
Transparent			Transparency and explainability
Understandable			

써 전세계의 개인정보보호 제도의 기준으로 인식되고 있는 것과 마찬가지로 인공지능 법안도 다른 국가들이 준비하고 있는 인공지능 법안의 기준으로 인식될 가능성이 크다. 유럽의 인공지능 법안은 시스템의 위험 등급을 4개로 나눠 규제 강도를 다르게 적용하는 접근 방식을 취하고 있다. 그 기준은 허용할 수 없는 위험, 고위험, 제한된 위험, 낮은 위험으로 분류하였는데 특히 허용할 수 없는 위험의 예로 개인에게 불리하게 작용할 수 있는 평가시스템의 이용 등을 들며 허용되지 않은 인공지능 시스템의 개발과 배포를 강제로 금지하고 있다는 점이 주목해볼 만하다. 이는 인간의 가치와 기본 권리를 엄격하게 평가 적용하고 있음을 보여주는 것으로 사람 중심의 인공지능 활용이 EU의 기본원칙임을 다시 한번 확인할 수 있다.

인공지능 기술의 신뢰성 및 윤리성을 담보하기 위한 노력은 국가 차원에서뿐만 아니라 국제기구 차원에서도 이뤄지고 있다. 2017년 IEEE(국제전기·전자공학자협회)에서 발표한 'Ethically Aligned Design, Version 2[30]'는 자율주행차, 인공지능 챗봇과 같은 자율적이고 지능적인 시스템(A/IS)의 개발과 사용에 대한 원칙을 제시하였다. 2019년 5월 22일 경제협력개발기구(OECD)는 '인공지능 윤리 원칙[31]'을 발표했으며 이후에 발표된 미국, 영국 등 주요국들의 인공지능 윤리 원칙 개발에 큰 영향을 주었다.

반면, 국내에서 공개된 인공지능 정책·제도의 특징을 살펴보면 미국과 유럽연합의 중간 정도로 아직 구체화된 모습을 보이지 않고 있다. 국내에서는 정부 주도로 인공지능 기술 개발과 혁신을 촉진하며 기술 규제를 최소화하고 기업의 자율성과 기술 혁신의 필요성을 강조하고 있으면서도 이와 동시에 인간의 가치와 개인정보 보호를 위해 중시한다는 점을 기본원칙을 내세우고 있다. 이를 토대로 인공지능 시스템의 안전성, 공정성, 투명성, 책임성 등을 확보하기 위한 윤리 원칙을 강조하고 있다. 예를 들어, 2019년 6월 정부 합동으로 발표한 '인공지능 윤리 원칙[32]'과 2021년 11월 발표한 '인공지능 윤리 가이드라인[33]'을 살펴보면 인공지능 기술의 개발과 혁신을 촉진하면서도, 인간의 가치와 개인정보를 보호해야 한다고 밝힌 부분이다. 이에 더해 2021년 6월 방송통신위원회에서 '디지털 미디어 플랫폼'이 제공하는 인공지능 기반 추천 서비스의 불투명성과 편향성 문제점을 지적한 적이 있으며[34] 2022년 5월 국가인권위원회는 인공지능의 개발과정에서 발생할 수 있는 인권 침해와 차별에 대한 위험성을 경고하며 '인공지능 개발·활용에 관한 인권 가이드라인[35]'을 만든 바 있다. 이런 가운데 2023년 2월 인공지능 기업 지원법(일명 인공지능산업 육성 및 신뢰 기반 조성 등에 관한 법률 제정안)[36]이 국회 상임위 법안소위에서 통과되었는데 이로써 국내에서도 법률적 제도화가 본격적으로 논의되게 되었다. 위의 정책 사례들을 볼 때 국내 인공지능 기술에 대한 규제 방향은 아직 명확하지 않다. 기술 혁신을 촉진하는 미국의 접근 방식과 인간의 사회적 가치를 중시하는 유럽연합의 접근 방식 사이에서 균형을 찾고 있다고 해석할 수 있다.

4.2 주요 인공지능 윤리 원칙 비교

인공지능의 기술 지배력을 두고 국가간 경쟁이 과열되고 있다. '인공지능 기술을 어떻게 효율적으로 통제할 수 있는냐'가 향후 각 국가들의 기술 혁신과 미래 가능성에 지대한 영향을 미치게 될 것이다. 인공지능 시스템이 사회에 긍정적인 영향을 미치기 위해서는 우선적으로 신뢰할 수 있어야 한다. 인공지능 시스템이 신뢰할 수 없다면 사용자는 인공지능 시스템을 사용하지 않을 것이며 낮은 안정성으로 인해 시스템이 악용할 가능성 또한 높아질 것이다. 따라서 인공지능 시스템을 개발하고 사용하는 모든 이해 관계자들은 인공지능 시스템의 신뢰성을 보장하기 위한 시스템을 만들고 이를 준수해야 할 것이다.

딥러닝 알고리즘의 등장으로 인공지능 기술이 빠르게 진화하게 되었고 생성형 인공지능을 대표로 하는 인공지능 기술의 급격한 사회 확산은 정책 입안자인 정부 당국과 사용자 모두를 당황하게 하였다. 이에 따라, 2019년부터 2021년까지 집중적으로 많은 국가와 국제기구들이 인공지능 윤리 원칙을 발표하였다. 그중에서도 대표적인 인공지능 원칙을 비교해 보면 <표 2>와 같다. 이들 원칙들은 전반적으로 유사한 주제와 가치를 강조하고 있지만, 경험하고 있는 고유의 사회문화적, 경제적, 정치적 환경에 따라 가치 기준이 다르며 윤리 원칙을 강조하는 방식에 있어서도 차이를 보이도 있다. 주요 국가들과 기업들이 강조하고 있는 윤리 원칙들을 보면 몇 가지 주요 공통점을 확인할 수 있는데 <표 3>과 같다.

주요 국가들과 기업들이 강조하는 윤리 원칙들에서 투명성(Transparency), 공정성(Fairness), 개인 정보 보호(privacy), 설명 가능성 (Explainability), 보안성·안전성(Security) 이 다섯 가지를 공통적으로 확인할 수 있는데 그 이유는 이 원칙들이 인공지능 기술이 사회 전반에 긍정적인 영향을 미치며 동시에 기술에 관한 부정적인 인식을 최소화하는 데 가장 중요한 핵심 요소이기 때문이다. '투명성'은 기술의 모호함을 없애고 기술의 합법성 및 적절성을 평가하기 위한 핵심 요소가 되며,

표 3. 인공지능 윤리 원칙의 공통 요소(특성)
Table 3. Common Property of AI Ethics

Property	Features and Meaning
투명성 (Transparency)	인공지능 시스템의 동작 방식과 의사결정 과정이 명확하게 이해될 수 있어야 한다는 원칙, 사람들이 인공지능 결과에 대해 신뢰를 갖는 데 필요한 중요 요소
공정성 (Fairness)	인공지능 시스템이 부당한 차별이나 편향 없이 모든 사람에게 공정하게 작용해야 한다는 원칙, 편향된 데이터로 인해 발생하는 차별적인 결과를 방지하는 것이 중요
개인 정보 보호 (Privacy)	인공지능 사용자의 개인정보 보호, 사생활 존중, 그리고 인간 가치 본연의 자율성과 연계되어 인권 자체와 밀접히 연관
설명 가능성 (Explainability)	복잡한 인공지능 시스템, 특히 심층 학습과 같은 기술은 작동 원리가 비공개적이거나 이해하기 어려움, 사용자들이 AI를 이해하려면, AI는 그 결정 과정을 설명할 수 있어야 하며 이는 투명성 원칙과 밀접히 관련
보안성·안전성 (Security)	인공지능 시스템이 예상치 못한 방식으로 행동하거나 공격당하지 않아야 함을 의미, 개인 정보 보호에 직접적인 영향을 미치며, 또한 사회 전반에 걸쳐 AI의 잠재적인 부정적인 영향을 방지하는 데 중요

'공정성'은 편향된 알고리즘으로 인한 사회적 불평등이나 차별을 최소화함으로써 기술에 대한 인간의 거부감을 해소해 준다. 또한 '개인정보보호'는 데이터 중심의 디지털 사회에서 인간의 가치와 기본 권리에 대한 기본 욕구이며, '설명 가능성'은 인공지능이 왜 이렇게 결정했는지를 설명해 줌으로써 그 결정을 이해하고 그 결정에 대해 이의 제기를 가능하게 해주며. 마지막으로 '보안성·안전성'은 인공지능 시스템이 해커 등으로 인해 악의적으로 조작되지 않음을 보장함으로서 기술에 대한 신뢰감을 부여해 준다.

4.3 국내 인공지능 정책 적용 방안

본 절에서는 앞 절에서 살펴본 해외 주요국들의 정책 내용을 토대로 국내에 적합한 법률 및 정책 방안, 신뢰성 ·투명성 기준 수립, 인공지능 기술 평가·인증체계 도입, 이해관계자 참여 확대, 글로벌 공조 및 표준화 활동 강화 등 국내 실정에 적합한 정부 정책 방안을 제시하고자 한다.

4.3.1 법률 및 규제 정책의 정비

인공지능 관련 법률 및 규제 정책의 정비는 인공지능과 관련된 사회적, 윤리적, 법적 문제를 효과적으로 해결하는 핵심 방법의 하나이다. 정부 정책이 기술 진흥에 지대한 영향을 미치고 있는 국내 실정에서는 더욱 그러하다. 기술의 사회적 확산에 앞서 관련된 법률 정비를 통해 다양한 사회적 이슈에 대응하며, 기술의 안전성과 신뢰성을 확보할 수 있어야 한다. 현재까지 제안된 여러 인공지능 관련 법안들은 인공지능 기술의 표준 관리 체계나 위험성에 대한 구체적인 대응 방안이 적시되어 있지 않고 선언적인 요소로 구성되어 있다. 특히 법안에서는 정부 및 민간 기업이 어떤 역할과 책임을 져야 하는지 명시되어야 한다. 인공지능 알고리즘 또는 시스템 자체를 대상으로 하는 구체적인 위협 또는 취약점이 확인되면 반드시 개선 조치가 이루어지도록 이행을 담보해야 하고 이행하지 않을 시의 처벌조항도 구체화하여 실효적인 법률체계로 정립해 나가야 한다. 또한. 법률 및 규제의 지속적 개선과 갱신을 통해 적절한 규제 환경을 조성하며, 다양한 이해관계자들의 폭넓은 의견 수렴과 다양한 시각을 반영해야 한다. 마지막으로, 인공지능 정책의 지속적인 모니터링 및 평가를 통해 현재의 기술 트렌드와 사회 변화에 맞게 제도 및 정책이 변화해야 한다. 궁극적으로는 국회에서 논의 중인 인공지능 관련 법안을 조속히 통합 검토하여 기술 발전에 따른 사회적 우려를 줄여 나가야 할 것이다.

4.3.2 인공지능 기술의 신뢰성·투명성 기준 마련

인공지능 기술의 사회적 적용 확대를 위해서는 필연적으로 그 사회적 파급영향과 이해관계자들의 우려를 고려하여야 한다. 기술 발전으로 인한 정보 불균형, 디지털 격차, 노동 일자리 변화 등의 사회적 문제를 해결하는 방안을 모색하고 정책에 반영할 필요가 있다. 인공지능 기술 알고리즘에 대한 신뢰성·투명성을 강조하여 인공지능 정책의 개발과정과 결과물을 투명하게 관리해야 한다. 유럽의 인공지능 법안에서와 같이 국내 도입·활용되는 인공지능 시스템의 위험도를 구분 분류하고, 위험의 발생 가능성, 발생시 예상되는 피해 영향 등을 포함하여 국가적인 표준으로 마련해야 한다. 이를 위해서는 기술 표준화 담당 기관(국가기술표준원 등), IT기술진흥 기관(정보화진흥원 등), 보안 전문기관(국가보안기술연구소) 등 다양한 분야의 전문가가 참여하여 예상되는 모든 위험을 다루어야 할 것이다. 이를 통해 인공지능 기술이 파생하고 있는 기술적 위험성에 대한 기본 대책을 마련하여 인공지능 기술의 안전성과 신뢰성 문제에 대한 사용자들의 우려를 불식시켜야 한다.

특히, 인공지능의 신뢰성·투명성 개념은 인공지능 작동 원리와 그 영향력이 이해할 수 있게 개발되어야 함을 의미하며 이를 위해서는 인공지능 시스템의 결과가 어떻게 도출되었는지에 대한 설명이 구체적으로 공개되어야 하며 인공지능의 결정 과정이 공정하게 이루어졌는지 확인할 수 있는 감시 제도가 별도로 마련되어야 한다. 마지막으로 정부가 앞서서 공공 데이터 플랫폼을 기반으로 인공지능 소스 코드와 학습 데이터 공개를 선도함으로써, 인공지능 혁신을 촉진하는 제도로 확립해야 한다. 이는 다른 연구자들이 알고리즘의 신뢰성을 검증하는 데 도움이 될 것이며 이를 통해 인공지능 기술에 대한 신뢰성과 투명성이 확보되어 사용자 신뢰를 증진시키게 된다.

4.3.3 인공지능 평가·인증체계의 도입

인공지능의 위험성을 초래하는 원인으로는 시스템 개발시의 기술적인 알고리즘의 결함, 학습 데이터의 부정확성, 해킹 등의 악의적인 공격 등 여러 요인을 들 수 있다. 따라서 인공지능 시스템의 신뢰성을 보장하기 위해서는 체계적인 평가 및 인증 시스템 구축이 고려되어야 한다. 이를 위해서는 첫째, 신뢰할 수 있는 인공지능 소프트웨어가 무엇을 의미하는지에 대한 표준 정의가 이루어져야 한다. 이어서 성능 지표, 보안 기능과 함께 공정성, 책임성, 투명성과 같은 윤리적 고려사항 등이 포함된 평가체계로 마련되어야 한다. 둘째, 인공지능 신뢰성을 평가하기 위한 평가기관(또는 시험기관)은 인공지능 소프트웨어가 설정된 표준을 충족하는지 평가하기 위한 엄격한 테스트 절차를 만들어 공개하고 이해 상충을 방지하기 위해 독립적인 제3자에 의해 그 평가가 수행되어야 한다. 이는 정부 기관, 비영리 조직, 또는 인증 과정의 결과에 이해관계가 없는 민간 회사가 될 수 있다. 셋째, 평가 기준과 인증 과정의 결과는 개발자, 사용자, 규제 당국 등 모든 이해관계자에게 투명하게 공개되어야 한다. 마지막으로 인공지능 소프트웨어는 시간이 지남에 따라 발전하고 학습하는 특성을 가지므로 지속적인 모니터링을 통해 적절한 인증 상태를 유지하고 있는지 업데이트 과정을 확인해야 한다.

4.3.4 다양한 이해관계자 참여 확대

인공지능 기술의 윤리적 측면에 대한 고려는 단순히 정부 차원의 정책 수립뿐만 아니라, 기업, 학계, 그리고 일반 시민들에게도 영향을 미친다. 이에 따라, 다양한 이해관계자들이 함께 참여하는 토론회, 워크숍, 세미나 등을 통해 인공지능 윤리에 대한 기본 인식을 높이고, 소통의 기회를 확장하는 것이 필요하다. 또한, 국민의 참여와 의견을 적극 수렴하며 인공지능 관련 정책을 수립하는 것이 중요하다. 시민들의 목소리와 눈높이에서 정책을 고려하고 실행함으로써, 기술의 발전이 사회 전반에 걸쳐 긍정적인 영향을 미칠 수 있도록 해야 한다. 이를 위해서는 인공지능과 관련된 다양한 교육 프로그램이 개발되어야 하며, 미래 인재들이 이 분야에서 성과를 이룰 수 있도록 지원되어야 하는데 기술적 역량 지원 외에도 윤리적인 가치를 함양하는 교육도 동시에 추진되어야 할 것이다.

4.3.5 글로벌 공조 및 표준화 활동

인공지능 기술은 이미 한 국가를 넘어서 국경을 초월하는 영향력을 가지게 되었다. 특히 핵심 기술들은 글로벌 기술 기업들이 주도하고 있는 현실에서 이를 관리하고 효과적으로 활용하기 위해서는 국가 간 협력과 글로벌 표준 수립이 필요적이다. 따라서, 글로벌 인공지능 거버넌스 구축을 위한 국제 협력을 강화하고 다양한 국가와 지역 간의 경험과 지식을 공유하는 것이 중요하다. 이를 통해 글로벌 차원에서 인공지능 기술의 발전과 윤리적인 문제에 대해 함께 고민함으로써 올바른 해결책을 찾는 노력이 병행되어야 한다. 이런 과정을 통해 인공지능 기술이 가져올 혜택은 더욱 확대하고, 동시에 윤리적인 측면과 안전성을 보장할 수 있다. UN, OECD, EU 등과 같은 국제적인 조직들은 이미 인공지능에 대한 다양한 가이드라인과 규정을 제안하고 있다. 이러한 국제기구들과의 협력을 통해 알고리즘의 투명성, 데이터 처리 방식 등에 대한 표준화 활동에도 적극 참여, 글로벌 인공지능 규제 표준을 만드는 데 구체적인 기여가 있어야 한다. 이를 통해 국내 인공지능 기술에 대한 글로벌 영향력을 확대함과 동시에 인류의 공통된 가치와 원칙을 확립하는 데 있어 한국이 중요한 역할을 한다는 인식을 만들어 나갈 수 있다.

V. 결 론

인공지능 기술의 혁신성이 전세계의 경제 산업계뿐만 아니라 사회 전 분야에 영향을 주고 있다. 최근 사회적 이슈가 되었던 ChatGPT를 대표로 하는 생성형 AI 기술들은 인간의 일하는 방식을 근본적으로 바꾸어 나가고 있다. 그러나 전례 없는 빠른 속도로 진화되는 기술은 미처 적응하지 못한 사회에 큰 우려 또한 만들고 있다. 초기 인공지능 개발을 이끌었다 학자들조차도 이제는 적절하게 통제되지 않은 기술이 불러올 수 있는 해를 경고하고 있다. 이런 상황에서 정부는 산업계, 학계 등 분야별 이해관계자들의 다양한 의견을 모아 기술

혁신과 함께 기술의 신뢰성을 동시에 균형 있게 끌어나가야 할 것이다.

그런 의미에서 본 연구에서는 인공지능 기술이 사회에 미치는 긍정적인 요인과 함께 부정적인 요인이 무엇인지 확인하고 그 파급력을 고려하여 올바른 정책 방향을 제시하고자 하였다. 그러기 위해서 우선 인공지능 기술 자체가 보유하고 있는 안전성 문제와 인공지능이 사회 각 영역에 적용됨으로써 생기게 될 대표적인 사회적 역기능과 그 위험성이 무엇인지 구분하여 살펴보았다. 이어 인공지능 기술이 사회 전반에 걸쳐 전면적으로 도입되면서 미국, 유럽 등 인공지능 기술을 선도하는 국가들이 인공지능 신뢰성을 확보하기 위해 다각적으로 도입하고 있는 정부 정책들과 함께 그 정책들의 변화상을 살펴보고 각 국가별 공통점과 차이점을 도출해 보았다. 각 나라의 인공지능 정책 분석을 통해 국내 정부 정책에 적용할 수 있는 시사점을 도출해 적용할 수 있다면, 기술의 발전과 신뢰성 이슈를 더욱 효과적으로 다룰 수 있기 때문이다.

국내에 적합한 성공적인 인공지능 정책을 구축하려면 여러 핵심 요소가 고려되어야 한다. 우선 국내 인공지능 기술을 활성화하고, 그 부작용을 최소화하려면 체계적인 법률 및 규제 정책의 정비가 필요하다. 또한, 인공지능 기술은 그 구조와 원리가 복잡하고 불투명한 경우가 많으므로, 투명성을 확보하는 것이 필수적이며 신뢰할 수 있는 인공지능 시스템을 위해 안전성을 입증할 수 있는 평가·인증체계를 갖추는 것이 중요하다. 또한, 다양한 이해관계자의 참여를 공고히 하는 것은 인공지능 기술 규제의 다양한 측면을 고려하게 하고, 기술 혁신을 촉진하는 중요한 매개체 역할을 한다. 마지막으로, 국내 외에도 국제적인 협력을 통해 기술의 신뢰성을 확보하는 데 필요한 글로벌 표준을 구축해 나가는 것이 중요하다. 인공지능 기술의 확산은 거스를 수 없는 시대적 조류가 되었으며 인공지능과 인간 사회의 공존을 위해서는 기술을 안전하게 사용할 수 있다는 기술의 신뢰성이 담보되어야 한다. 이 같은 전제가 선행되어야만 인공지능 산업 경쟁력과 함께 사이버공간의 안전성이 확보되고 더 나아가 한 국가의 사이버안보 또한 확고히 유지하게 될 것이다.

References

[1] S. Oh, et al., "Strategies for improving STI policy towards a leading country in the application of ai technology," *A policy study*, pp. 1-404, 2020.

[2] J. Yoo, "Domestic and foreign policy trends to realize reliable AI," *KISDI AI Outlook*, vol. 6, pp. 17-32, 2021.

[3] H. Jang, "A study on policy instrument for the development of ethical ai-based services for enterprises: An exploratory analysis using AHP," *JITS*, pp. 23-40, 2022. (https://doi.org/10.22693/NIAIP.2022.29.1.060)

[4] G. Kim "Explanable artificial intelligence technology trends to secure the reliability of artificial intelligence systems," *Korea Inf. Process. Soc. Rev.*, vol. 28, no. 3, p. 45, Sep. 2021.

[5] K. Kim, "Artificial intelligence trustworthiness international standard trend for industrial digital transformation," *Conf. Korean Soc. Electron. Eng.*, vol. 2022, no. 06, pp. 2674-2678, 2022.

[6] H. yang, "Safety issues of ai and policy responses," *J. KICS*, vol. 43, no. 10, 2018. (https://doi.org/10.7840/kics.2018.43.10.1724)

[7] S. Jung, "European parliament resolution on the use of artificial intelligence in criminal justice - Regarding the ban of automated decisions by AI," *Korean J. Criminology*, vol. 34, no. 2, Jul. 2022.

[8] H. Kim, "Artificial intelligence ethics and criminal policy - Focusing on the issue of artificial intelligence ethics in criminal justice -," *Korean J. Criminology*, vol. 34, no. 1, Apr. 2022.

[9] H. Kim, "Ethics-based control of high-risk ai technology in criminal law and policy," *Korean J. Criminology*, vol. 34, no. 1, Apr. 2022.

[10] S. park, "AI security issues," *J. Korea Inst. Inf. Secur. & Cryptol.*, vol. 27, no. 3, pp. 27-32, Jun. 2017.

[11] S. Lee, "Machine learning model attack research trends: Focusing on deep neural networks," *J. KIISC*, vol. 29, no. 6, pp. 67-74, Dec. 2019.

[12] MBN TV, "*Preventing the "Second Luda Incident"...LG and Kakao,*" Retrieved Jul. 12, 2023, from http://mbnmoney.mbn.co.kr/news/

view?news_no=MM1004717257

[13] Y. Shen, "Reinforcement learning algorithm based hybrid filtering image recommender system," *The J. Inst. Internet, Broadcasting and Commun.*, vol. 12, no. 3, pp. 75-81, 2012. (http://dx.doi.orgI10.7236/J1WIT.2012.12.3.75)

[14] NIST of U.S., "*Demographics study on face recognition algorithms*," Retrieved Jul. 12, 2023, from https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects.

[15] AI Times, "*The 'Escape Prompt'*," Retrieved Jul. 12, 2023, from https://www.aitimes.com/news/articleView.html?idxno=150437

[16] ZDNet "*I feel a heavy responsibility for personal information*," Retrieved Jul. 12, 2023, from https://zdnet.co.kr/view/?no=20210428195323

[17] E.U. Commission, "Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment," 17 Jul. 2020, Retrieved Jul. 12 from https://digitalstrategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

[18] Digital Service of U.K., "Data ethics framework," 13 Jun. 2018. (Updated 16 Sep. 2020), Retrieved Jul. 12, 2023, from https://www.gov.uk/government/publications/data-ethics-framework

[19] Department for Science, Innovation & Technology Policy of U.K., "National data strategy," 9 Sep. 2020, Retrieved Jul. 12 2023, from https://www.gov.uk/government/publications/uk-national-data- strategy

[20] GAO of U.S., "Artificial intelligence: An accountability framework for federal agencies and other entities," Jun. 30, 2021, Retrieved Jul. 12 2023, from https://www.gao.gov/products/gao-21-519sp

[21] NIST in U.S., "AI risk management framework," Jan. 26, 2023, Retrieved Jul. 12 2023, from https://www.nist.gov/itl/ai-risk-management-framework

[22] White_House of U.S., "American AI initiative," Feb. 10, 2019, Retrieved Jul. 12 from https://trumpwhitehouse.archives.gov/ai/

[23] U.S. Congress, "National AI initiative act of 2020," Jan. 1, 2021, Retrieved Jul. 12 2023, from https://www.congress.gov/116/crpt/hrpt617/CRPT-116hrpt617

[24] NIST of U.S., "Technical AI standards," Aug. 3, 2021, Retrieved Jul. 12 from https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf

[25] E.U. Commission, "Artificial intelligence for Europe (EU AI Strategy)," 2018.4.25, Retrieved Jul. 12 2023, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN

[26] E.U. Commission, "Ethics guidelines for trustworthy AI," 2019.4. 8, Retrieved Jul. 12 2023, from https://digital‐strategy.ec.europa.eu/en/library/ethics-guidelines

[27] Department of Industry, Science and Resources in Australia, "*Australia's Artificial Intelligence Ethics Framework*," 7 Nov. 2019, Retrieved Jul. 12 2023, from https://www.industry.gov.au/publi-cations/australias-artificial-intelligence-ethics- framework

[28] Department for Digital, Culture, Media & Sport of U.K., "National data strategy," 9 Dec. 2020, Retrieved Jul. 12 from https://www.gov.uk/government/publications/uk-national-data-strategy/national-data

[29] E.U. Commission, "The AI Act," 21.4. 2021, Retrieved Jul. 12 2023, from https://artificialintelligenceact.eu/the-act/

[30] IEEE, "Ethically aligned design ver 2," Dec. 2017, Retrieved Jul. 12 2023, from https://standards.ieee.org/industry-connections/ec/ead-v1/

[31] OECD, "OECD Principles on AI," 2 Mar. 2020, Retrieved Jul. 12 2023, from https://www.oecd-ilibrary.org/economics/what-are-the-oecd-principles-on-ai_6

[32] Relevant ministries of the Republic of Korea, "Artificial intelligence national strategy," 2019. 12. 17, Retrieved Jul. 12, 2023, from https://www.msit.go.kr/bbs/view.do?sCode=user&mId=113&mPid=112&pageIndex=1

[33] Relevant ministries of Korea, "AI ethical

standards," 2020.12.23, Retrieved Jul. 12, 2023, from https://www.msit.go.kr/bbs/view.do?sCode=user&nttSeqNo=3179630&pageIndex=&searchTxt=&searchOpt=ALL&bbsSeqNo=94&mId=113&mPid=112

[34] Korea Communications Commission, "Basic principles for user protection of ai-based media recommendation service," Jun. 30, 2021, Retrieved Jul. 12, 2023, from https://kcc.go.kr/user.do?boardId=1113&page=A05030000&dc=K00000200&boardSeq=51454&mode=view

[35] National Human Rights Commission of Korea, "Human rights guidelines on AI," May 17, 2022, Retrieved Jul. 12, 2023, from https://www.humanrights.go.kr/site/program/board/basicboard/view?boardtypeid=24&boardid=7607961&menuid=001004002001

[36] Maeil Business News, "Public intelligence act enactment, passing the national assembly bill subcommittee," *Maeil Economy*, 2023-02-14. Retrieved Jul. 12, 2023, from https://www.mk.co.kr/news/it/1064 4221

[37] U.S. Chamber, "Artificial Intelligence Principles," Sep. 23, 2019, Retrieved Jul. 12 2023, from https://www.uschamber.com/assets/archived/images/chamber_ai_principles_-_general.pdf

[38] White_House, "Promoting the Use of Trustworthy Artificial Intelligence in Government," Dec. 3, 2020, Retrieved Feb. 21 2023, from https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence

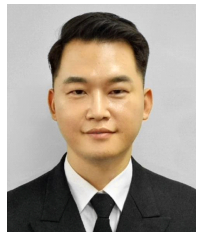
[39] Department of Industry, Science and Resources of U.K., "Understanding AI ethics and safety," 10 Jun. 2019, Retrieved Jul. 21 from https://www.gov.uk/guidance/

김 진 민 (Jinmin Kim)

2000년 2월 : 숭실대학교 컴퓨터학부 졸업
2002년 2월 : 포항공과대학교 컴퓨터공학과 석사
2022년 3월~현재 : 숭실대학교 IT정책경영학과 박사과정

<관심분야> 사이버안보, IT보안정책, 정보보안

이 민 철 (Mincheol Lee)



2011년 3월 : 해군사관학교 외국어학과 졸업
2018년 6월 : University of Delaware Electrical and Computer Engineering 석사
2022년 3월~현재 : 숭실대학교 IT정책학과 박사과정

<관심분야> 전자공학, 통신공학, 정보보호

서 정 훈 (Junghun Seo)

2022년 2월 : 숭실대학교 IT정책 경영학과 석사
2022년 3월~현재 : 숭실대학교 IT정책경영학과 박사과정
현재 : Altibase(알티베이스) 사업 본부 근무 중

<관심분야> AI, 금융공학, 분산·메모리 DBMS

신 용 태 (Yongtae Shin)

1985년 2월 : 한양대학교 산업공학과 (공학사)
1990년 12월 : Univ. of Iow, Computer Science (공학석사)
1994년 5월 : Univ. of Iow, Computer Science (공학박사)
1995년 3월~현재 : 숭실대학교 컴퓨터공학부 교수

<관심분야> 컴퓨터 네트워크, 정보보호 기술.