영국의 이용자 보호를 위한 AI 윤리 지침: 'Understanding Artificial Intelligence Ethics and Safety'를 중심으로

변 성 혁*

1. 개요

최근 인공지능(Artificial Intelligence, 이하 AI) 기술이 발달하고 AI 기반의 시스템이 고도화됨에 따라 그 활용 영역이 날로 증가하고 있다. 우리나라를 비롯한 영국, 미국 등 주요국은 국가 발전의 진흥을 도모하고 사회 여러 분야 전반의 성장을 위해 AI와 빅데이터, IoT 등 4차산업혁명 기술에 전(全) 국가적 역량을 집중하고 있다. 이로 인해 우리 사회 곳곳에 생산성과 효율성이 제고되는 등 국가 전반에 걸쳐 긍정적 영향이 나타나고 있다. 그러나 이러한 AI 기술이 반드시 이로운 것만은 아니다. 빅데이터와 개인정보가 수집되고 활용되면서 사회 구성원 개개인의 프라이버시 침해 문제가 발생하고 있으며 디지털 취약계층에게 불리한 알고리즘으로 인해 정보의 빈익빈부익부 현상이 더욱 심화되고 있기 때문이다. 이는 이용자의 편향(bias)과 디지털 디바이드(Digital divide) 등 또 다른 사회적 문제를 발생시키며 역기능을 초래하는 결과로 이어질 수 있다.

이러한 부작용에 선제적으로 대응하기 위해 우리나라를 포함한 세계 주요 국가에서는 이용자 보호를 위한 AI 윤리 원칙을 연구하고 논의를 진행하고 있다. AI 윤리 관련한 이용자 보호 이슈를 도출하고 사회 각계각층의 전문가들이 참여하는 정책연구를 수행하며, AI

* ICT전략연구실 연구원, 043-531-4398, shbyun28@kisdi.re.kr

시스템으로 인해 발생 가능한 잠재적 위험을 선제적으로 식별하기 위해 노력하고 있다. 이러한 활동은 ICT 관련 기술을 이용하는 사회 구성원들의 편익을 높임으로써 순기능을 최대화하고, 알고리즘 편향과 프라이버시 침해로 인한 불평등과 차별, 사생활 침해 등 부정적인 역기능을 최소화하기 위해 상당히 중요한 것이라 할 수 있다.

그 중, 영국은 AI 윤리 관련 연구를 선도하는 대표적인 국가로서 세계 최초로 데이터윤리혁신센터[1]를 설치하였으며 앨런 튜링 연구소(Alan Turing Institute)[2]와 협력하여 안전한 AI 활용에 관한 지침을 개발하는 등 공정한 AI 활용을 위해 노력해왔다. 특히, 최근 앨런 튜링 연구소에서 개발한 AI 이용자 보호 윤리에 관한 지침은 AI로 인해 발생할 수 있는 역기능과 잠재적인 위험을 발견하고, 이를 조기에 방지하기 위한 구체적인 조치를 제시하여 유의미한 정보를 제공하고 있다. 이에 본고에서는 AI 윤리의 개념 및 중요성을 간략히 정리하고, 앨런 튜링 연구소에서 발간한 'Understanding Artificial Intelligence Ethics and Safety(인공지능 윤리와 안전의 이해, 2019.06)' 지침서의 주요 내용을 소개하도록 하겠다.

2. AI 윤리의 개념 및 중요성

AI 윤리란, AI 시스템을 개발하고 활용할 때 허용되는 도덕적·윤리적 행동의 기준과 원칙을 의미한다. 즉, 효율성과 생산성 극대화 측면에만 몰두하여 AI 시스템을 개발하는 것이 아닌, 잘못된 알고리즘 설계, 데이터 편향 등과 같은 부정적 영향을 최소화하고 안전성과 윤리적 측면 등 사회적 공익에 부합할 수 있도록 설계하는 것을 의미한다. 특히, 본

1) 인공지능 관련 윤리적이고 혁신적인 데이터를 사용하기 위한 최선의 연구 및 실행방안을 마련하고, 구체적인 정책이나 관련 규제가 필요할 경우 정부에 자문하는 기능을 수행하는 기관

2) 데이터 과학과 인공지능에 대해 연구하는 기관으로서, 2015년 케임브리지, 에든버러, 옥스퍼드 등 5개 대학과 영국의 공학과 물리과학 연구위원회가 설립. 본 연구소는 수학, 통계학, 컴퓨터 과학, 사회 과학 및 데이터 윤리학, 소프트웨어 공학, 기계 학습 및 인공지능 등 다양한 분야의 연구원들을 모아 데이터 과학 분야에서 세계 최고 수준의 연구를 수행하고 있음

지침서에는 AI 윤리를 'AI 시스템 설계 시, 사회적으로 허용 가능한 도덕적·윤리적 가치, 원칙, 기법들의 집합체'라고 정의하며, AI로 인한 개인 및 사회적 피해를 최소화해야 함을 강조하고 있다.

AI로 인해 발생할 수 있는 주요 역기능의 내용은 다음과 같다. 먼저, 편향된 알고리즘과 차별(Bias and Discrimination)의 문제가 발생할 수 있다. AI는 시스템 개발자에 의해 설계되기 때문에, 잠재적으로 개발자의 선입견과 편견을 그대로 반영하여 편향될 수 있다. 또한, 우리 사회의 현 구조와 현상을 있는 그대로 학습하기 때문에 기성 사회의 불평등, 차별의 양상을 재현 혹은 증가시킬 수 있다. 두 번째로, AI 시스템으로 인한 피해 발생 시 책임 소재가 불분명 할 수 있다. AI 시스템의 설계 과정이 분산화 되어있고 구현 과정이 복잡하여, 알고리즘을 통한 부정적 결과가 초래되었을 때 시스템 문제의 책임자를 특정하여 파악하기 어렵기 때문이다. 세 번째로, 사생활 침해 문제가 있다. AI 시스템을 설계하고 구현할 때, 이용자의 개인정보를 이용하게 되는데 이러한 데이터가 정보 주체의 동의를 얻지 못하고 활용되거나 유출될 위험이 있다. 마지막으로, 필터 버블(filter bubble)[3] 현상이 발생할 수 있다. AI를 활용한 지나친 개인화 시스템으로 인해, 자신이 선호하는 영역만 굳건히 확립된 채 다양한 사람과의 상호작용을 줄이고, 다른 세계관에 대한 노출을 제한함으로써 사회관계를 양극화시킬 수 있다.

AI 윤리의 안전하고 공정한 기준과 원칙을 바탕으로 이러한 역기능을 통제하지 못한다면, 무책임하고 부주의한 AI 시스템 설계 프로세스가 구현됨으로써 비윤리적인 AI 시스템이 우리 사회에 만연해질 수 있다. 이러한 결과는 개인과 사회 구성원 전체의 행복과 복지(well-being)등 공익 측면에 직·간접적인 피해를 줄 수 있으며, 사회적으로 유익한 AI 기술에 대한 대중의 신뢰를 떨어뜨릴 수 있다. 따라서, AI 시스템을 설계하고 활용할 때, AI 윤리의 역할이 상당히 중요한 부분이라 할 수 있다.

3) 이용자의 개인정보와 활동데이터 등에 기반하여, AI가 이용자가 선호할 만한 결과만을 제공. 그 결과 이용자의 관점을 벗어나는 다양한 정보로부터 분리되고 문화적·이념적 거품(bubble)에 갇히는 현상 발생

3. 'Understanding Artificial Intelligence Ethics and Safety' 지침서 소개

'Understanding Artificial Intelligence Ethics and Safety' 지침서는 데이터 사이언티스트와 AI 프로젝트 책임자, 개발자 등 AI와 관련된 사람들을 대상으로, 안전하고 공정한 'AI 윤리 기반 프로젝트'의 가이드라인과 'AI의 윤리적 활용을 위한 구체적인 지침'을 제시하였다.

(1) AI 윤리 기반 프로젝트의 주요 내용

본 지침서에서는 윤리적이며 안전한 AI 시스템 원칙을 구축하고 그 가치를 실현하기 위한 AI 윤리 기반 프로젝트의 가이드라인을 제공하며 '책임감 있는 AI 혁신문화'를 강조하고 있다. 책임감을 기반으로 한 AI 시스템의 혁신적인 문화를 실현하기 위해 다음과 같은 4가지 구체적 목표를 제시하였다.

〈표 1〉 윤리적이며 안전한 AI 가치 실현을 위한 목표 설정

NO	목표	내용
1	윤리적 허용 가능 범위 내 시스템 설계 (Ethically permissible)	• 윤리적 허용 가능 범위 내에서, AI 프로젝트가 주요 이해관계자 및 커뮤니티에 끼치는 긍정적 영향 보장
2	공정성 및 비차별성 시스템 설계 (Fair and Non-discriminatory)	• AI 프로젝트가 이해관계자 및 커뮤니티에 차별적으로 끼치는 부정적 영향을 인지하고, 이를 최소화하기 위해 노력 • 특정 개인 혹은 집단에게, 차별적으로 편향된 이익이 최소화 될 수 있도록 시스템 설계
3	신뢰 확보 시스템 설계 (Worthy of public trust)	• AI 시스템을 통해 산출된 최종 결과물에 대한 대중의 신뢰(견고성, 안전성, 신뢰성, 보안성) 확보
4	정당성 확보 시스템 설계 (Justifiable)	• AI 시스템 설계 및 구현 과정에서, 투명하고 해석 가능한 근거를 기반으로 윤리적 정당성 확보

자료: Alan Turing Institute(2019)을 바탕으로 재작성

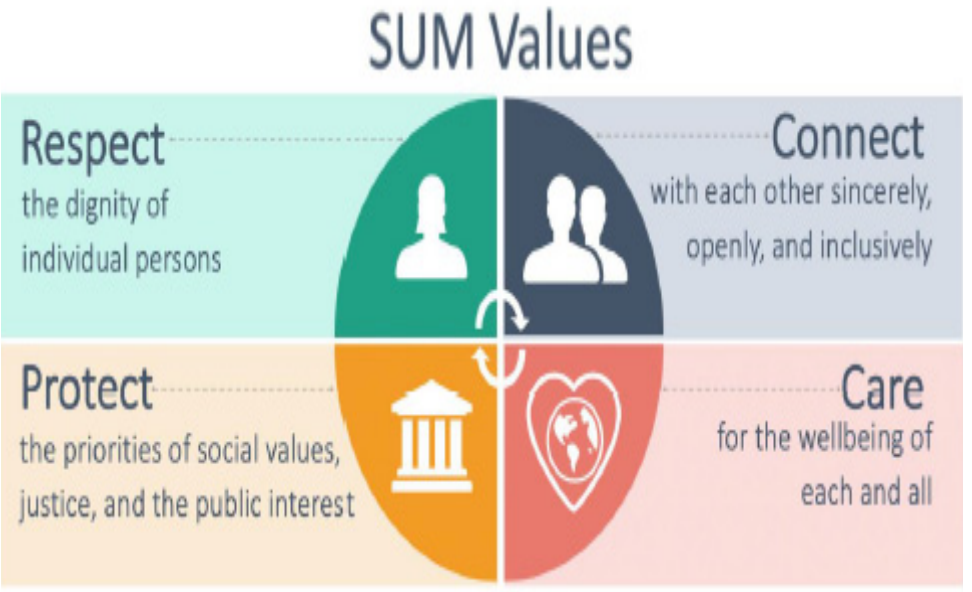
(2) AI의 윤리적 활용을 위한 지침

위와 같은 'AI 혁신의 책임감'을 실현하고 그에 따른 4가지 목표를 실행하기 위해, 본 지침서는 두 가지 프레임워크와 한 가지 원칙을 구체적으로 제시하였다. 보다 자세히, 윤리적 가치 프레임워크를 실현하기 위해 SUM 가치를, 실행 가능한 원칙 설정을 위해 FAST 원칙을, 프로세스 기반의 관리 체계를 구축하기 위해 PBG 프레임워크를 제안하였다.

1) SUM 가치

SUM 가치는 AI의 윤리적이며 안전한 시스템 원칙의 구축을 위한 첫 단계이다. SUM은 'Support, Underwrite, Motivate'의 약자로써, AI 기술과 데이터를 활용한 시스템 설계 시 모두가 지지하고(Support) 동의하며(Underwrite) 동기부여(Motivate)가 될 수 있는 가장 기초적인 윤리적 가치의 프레임워크를 제안한다. 즉, 책임감 있는 윤리 원칙을 목표로, AI 시스템을 설계하고 그것이 구현되었을 때 우리 사회와 개인에게 미치는 영향이 윤리적으로 허용 범위 내에 있을 수 있도록 프레임워크를 제공하는 것이다. 이때, SUM 가치를 실현하기 위한 구체적 요소로 Respect, Connect, Care, Project 총 4가지를 제시하였다. 개개인의 존엄함을 존중(respect)하며, 개개인 상호간에 진실되고 솔직한 연결(connect)을 통해 사회적 가치와 정의, 공익을 보호(protect)하고 사회 모든 구성원의 행복과 복지를 살필 때(care), SUM 가치가 실현될 수 있다고 보았다. 이러한 윤리적 프레임워크는 모두가 동의하고 인정하는 AI 시스템의 허용 가능한 윤리 요소를 제시하고, 그 기준을 설정하는데 의의가 있다고 할 수 있다.

[그림 1] SUM Values 구성요소



자료: Alan Turing Institute(2019)

2) FAST 트랙 원칙

FAST는 'Fairness, Accountability, Sustainability, Transparency'의 약자로써, AI 시스템의 탄탄한 설계와 활용을 위한 실행 가능한 실용적 원칙을 의미한다. SUM 가치는 윤리적 허용 가능성을 고려하는 원론적인 프레임워크를 제공하는데 의의가 있었다. 그러나 실용적 측면에서 AI 시스템을 설계하고 개발 및 실행을 하는데 있어서는 부적합할 수 있으므로 이를 보완하고자 'FAST 트랙 원칙'을 제시하였다. 이 원칙은 AI 시스템이 이용자에게 편파적이지 않고 차별적이지 않으며 공정하고 신뢰할 수 있는, 실질적인 실천 방안을 제공하는 것이 목적이다. 또한, 안전하고 믿을 수 있는 AI 혁신에 대한 믿음을 제공하기 위한 도덕적 도구를 제공하고자 하였다.

〈표 2〉 FAST 트랙 원칙

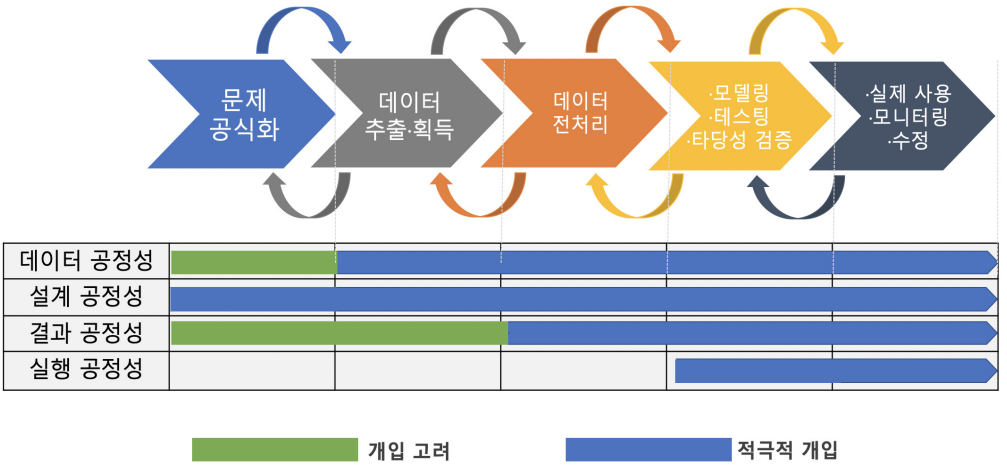
NO	목표	내용
1	공정성 (Fairness)	• 데이터 공정성: 공정한 알고리즘을 통한 책임 있는 데이터 수집 • 시스템 설계의 공정성: 비차별성을 바탕으로 한 시스템 설계 • 결과의 공정성: AI 시스템의 영향과 결과가 차별적이지 않아야 함 • 시행의 공정성: 베타단계의 AI 시스템의 공정한 구현 및 시행
2	책임 (Accountability)	• AI 시스템으로 인한 결과에 대한 책임 • AI 시스템 설계 및 구현 과정에 대한 책임
3	지속가능성 (Sustainability)	• 개인, 사회에 미칠 수 있는 지속가능한 변화와 장기적 영향력 인식 • 지속가능성의 요소: 정확성, 신뢰성, 보안성, 견고성
4	투명성 (Transparency)	• AI 시스템 설계 및 구현의 투명한 과정과 설명력 확보 • AI 시스템 결과의 윤리적 정당성 확보

자료: Alan Turing Institute(2019)을 바탕으로 재작성

3) PBG 프레임워크

PBG는 'Process Based Governance'의 약자로써, SUM의 원론적 가치를 적용시키고 FAST 트랙의 실용적 원칙을 작동하게 하는 프로세스 기반의 관리 프레임워크 구축을 의미한다. PBG 프레임워크의 목적은 AI 프로젝트의 정당성을 확보하고 투명한 시스템을 설계함으로써, 안전하고 윤리적인 AI 프레임을 구축하고 실현하는 것이다. 이를 위해, PBG 프레임워크는 각 단계별 구성원의 역할과 책임에 대한 정보를 구체적으로 제공하며, 거버넌스 목표를 달성하기 위해 할당된 각 단계별 역할 및 업무가 잘 수행·순환 되고 있는지 평가한다. 활용될 AI 시스템의 문제 발생 가능성 및 이슈를 검토하고, 이를 해결하기 위한 조치를 시행하는 것이다. 이에 따라, 필요한 경우 프레임워크에 대한 후속 조치 및 재평가를 진행하고 지속적인 모니터링을 통해 문제 해결에 필요한 과정까지 구체적으로 제공한다. PBG 프레임워크는 이처럼 명확한 워크플로(Workflow)를 바탕으로 AI 시스템의 기록과 감시를 지속적으로 관리한다.

[그림 2] AI 프로젝트 워크플로(Workflow)



자료: Alan Turing Institute(2019)을 바탕으로 재작성

4. 결어

영국은 세계에서 AI 분야를 선도하는 국가 중 하나이며, 공공분야에 AI 시스템을 활발히 적용하고 있다. 특히, AI 기술의 단순한 양적·질적 성장뿐만 아니라 사회 곳곳에 안전하고 윤리적인 AI 시스템을 적용하고 확립하기 위해 노력하고 있다. 이러한 가운데, 최근 영국의 앨런 튜링 연구소(Alan Turing Institute)는 'Understanding Artificial Intelligence Ethics and Safety' 라는 새로운 지침서를 발간하였고, 본 지침서가 AI 기술을 채택하고 활용하기 위한 올바른 가치와 윤리, 테크닉의 근간이 될 것이라고 밝혔다.

본 지침서에서는 AI 시스템이, 우리 사회에 윤리적으로 안전하며 책임 있게 개발되기 위한 구체적인 가치(SUM 가치)와 실용적인 원칙(FAST 트랙 원칙) 및 프레임워크(PBG 프레임워크)를 세부적으로 제시하였다. 이러한 지침은 기존의 여러 나라와 단체에서 제시한 원론적인 수준의 윤리 원칙·강령을 넘어서, 보다 상세하게 윤리적 수용 가능한 범위와 방법을 제공함으로써 다른 연구 보고서와 차별성이 있다고 할 수 있다. 또한, 이를 바탕으

로 윤리적이고 안전한 AI 시스템을 만드는 데 필요한 기본 의무를 규정하고 구체화함으로써, AI 기술의 위험성을 해결할 수 있는 지침서가 된다는 것에 중요한 시사점이 있다. 더불어, 본 지침서는 이전의 영국에서 발표한 데이터프레임워크[4]를 보완하였으며 AI 윤리를 확립하기 위한 도구로써의 가이드라인을 추가적으로 발전시켜 제시한 것에 그 의의가 있다고 할 수 있다.

우리나라에서도 100대 국정과제의 세부 실천과제로 '인공지능 윤리가이드(2017)'와 '인공지능 윤리헌장(2018)'을 발표하였으며, 윤리 가이드라인으로 공공성·책무성·통제성·투명성이라는 공통원칙을 수립하고, AI 기술의 주체별(개발자·공급자·이용자) 세부지침을 도출하였다. 그러나 현 시점에서 윤리적 이슈는 규범적 차원의 원론적인 논의 수준이며, 기술 발전의 양상과 빠르게 변화는 속도를 반영할 필요성이 대두되고 있다. 즉, 오늘날과 미래에 AI가 적용된 구체적 산업·서비스 모델을 감안하여 더욱 실효적으로 규범화할 수 있는 방법을 찾아가는 것이 필요하다.

이러한 상황을 종합적으로 고려해보았을 때, 'Understanding Artificial Intelligence Ethics and Safety'는 우리에게 원론적인 지침뿐만 아니라 실용적인 AI 윤리에 관한 실질적 지침을 제시한 것이라 할 수 있으며, 이를 한국사회와 기술 변화 양상에 맞게 적용한다면 우리의 산업과 연구 활동에 적용 가능한 유익한 지침서가 될 수 있을 것으로 보인다.

참고문헌

Alan Turing Institute(2019), "Understanding artificial intelligence ethics and safety", https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf

4) 정책 및 의사결정 과정에서, 데이터 사용이 증가함에 따라 고려해야 할 사용자 이익과 개인정보보호 및 보안 등에 관한 법적 측면과 원칙 등을 설명 (https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework)