AI 공정성을 위한 성숙도 모델 연구

(A Study on a Maturity Model for AI Fairness)

목 차

1. 서론

빅 데이터, IoT와 더불어 4차 산업혁명의 핵심기술인 AI (Artificial Intelligence)가 금융, 의료, 로봇, 자율주행 등 다양한 분야에서 널리 활용되고 있다. AI 영향력을 연구하는 전문가 모임인 AI100에서는 AI로 인해 인류의 삶이 2030년경 전방위로 바뀔 것으로 예측하고[1], IDC (International Data Corporation)는 글로벌 AI 시장이 2021년 3275억 달러에서 2024년에는 5543억 달러로 연 17% 이상 급속히 성장할 것으로 예상하였다[2].

AI를 구성하는 소프트웨어 (Software, 이하 S/W)는, 사람이 직접 구현한 규칙대로 동작하는 기존의 전통적 S/W와는 달리, 학습에 활용된 데이터에 따라 동작 방식이 결정된다[3]. 즉, 구현된 규칙에 따라 같은 상황에서 같은 동작을 보장하는 전통적 S/W와는 다르게, 학습된 데이터가 달라지면 같은 상황에서도 다른 동작을 할 수 있다. AI는 이러한 데이터 의존성으로 인해 사전에 예상하기 어려운 다양한 문제에 직면하게 된다. 예를 들어, 학습 데이터에 의도하지 않은 편향이 내재된 경우, 성별, 인종, 연령, 지역 차별과 같은 윤리 문제가 나타날 수 있다. 하버드 비즈니스 리뷰에서는 이러한 문제가 기업 브랜드 가치 하락의 원인이 될 수 있음을 경고하였다[4]. 다음은 최근 주목받았던 AI의 실제 문제 사례를 보여준다.

- 마이크로소프트 챗봇(Tay)은 부적절한 학습 데이터에 의해 욕설, 인종 및 성차별 발언(예, "히틀러가 옳아. 유대인은 싫어")을 하여 운영이 중단됨[5]
- 아마존의 구직자 평가 시스템에서는 남성 지원자가 많았던 과거 10년간의 채용 데이터에 의해 "여성"이 언급된 지원서를 채용 대상에서 배제하는 편향 문제가 발생함[6]
- 애플의 애플카드 알고리즘은 남성 위주의 데이터를 학습하여, 부부가 모든 자산과 계좌를 공유하는 애플 창업자 아내의 경우에도 남편 대비 신용한도가 10배 이상 낮게 평가되는 차별이 발생함[7]
- 구글의 비전인식 알고리즘에서는 체온계를 들고 있는 백인 손을 흑인 손으로 변경했더니, 총으로 오인하는 사례가 발생함[8]

위와 같은 사회적 이슈에 대응하기 위해서 각국 정부, IT 기업 및 국제기구는 AI 기술 개발과 사용에 대한 윤리 원칙을 공표해왔다. 구글은 불공정한 편향 방지, 안전성 확보를 염두에 둔 개발과 실험, 프라이버시 원칙 적용 등 7가지 AI 윤리 원칙을 발표했고[9], 마이크로소프트도 공정성, 신뢰성 등 책임 있는 AI 기본 원칙 6개를 정의하여 발표했다[10]. 또한, AI 윤리 관련 세계 최대 비영리 협회인 PAI (Partnership on AI)에서는 윤리적인 AI 연구 및 개발을 위한 6개 분야를 제시하였다[11].

이와 함께, AI 신뢰를 확보하기 위한 AI 국제 규제 및 인증 표준도 논의되었다. EU 집행위원회는 2021년 AI 신뢰 확보를 위한 AI 규제안을 발의하여, AI가 초래할 수 있는 위험 단계를 금지, 고위험, 제한된 위험, 최소 위험으로 구분하고, 고위험 AI 시스템을 공급하는 기업에게는 엄격한 요건 준수를 요구하고 있다[12]. 국내에서는 최근 한국표준협회가 AI 품질 인증 표준인 AI+ 인증을 공개했다[13].

추가적으로, 주요 IT 기업 및 대학에서는 편향 완화 기법, AI 시스템 개발 능력 및 프로세스 평가 등에 대한 연구를 수행하고 있다[14,15,16,17]. 다만, 아직 초기 단계로, 다양한 범주의 AI 공정성 이슈를 사전에 방지하기 위한 구체적이고 실행 가능한 정보를 제공하기에는 많은 연구가 요구된다.

본 연구에서는 공정하고 신뢰할 수 있는 AI S/W 품질을 확보하기 위한 AI 공정성 성숙도 모델인 AI-FMM (AI Fairness Maturity Model)을 제안한다. 이를 위해, S/W 개발 프로세스의 성숙도 모델인 SPICE (ISO/IEC 15504)를 기반으로, 체계적인 성숙도 모델을 개발하기 위한 방법론(Becker 등[18])을 적용하고, 기존의 다양한 연구 사례를 분석한다. 또한, AI 도메인 특성을 반영하기 위해 AI 전문가 인터뷰를 수행하고, 실제 운영 중인 다양한 과제에 적용함으로써 성숙도 모델의 실효성을 평가한다.

본 연구의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 AI S/W 특징, 공정성 및 편향, 전통적 성숙도 모델, AI S/W 프로세스 평가 등 관련 연구 동향을 분석하고, 3장에서는 공정성 이슈를 사전에 방지하기 위한 AI 공정성 성숙도 모델을 제안한다. 4장에서는 본 성숙도 모델의 실효성을 확인하기 위해 국내 IT 기업내 사례 연구 결과를 분석하고, 5장에서는 추가로 논의가 필요한 토론 항목을 제시하고, 마지막으로 6장에서 결론 및 향후 연구 방향을 도출한다.

2. 관련 연구

공정성에 영향을 미치는 요소로써 데이터가 한쪽으로 치우치게 되어 발생하는 편향과 이러한 편향을 개발 프로세스 상에서 방지하고 탐지하기 위한 개발 투명성에 대해 관련 연구를 살펴본다. 또한, 본 연구에서 제시하는 성숙도 모델의 기반이 되는 SPICE와 이를 확장하기 위한 성숙도 모델 개발 방법론에 대해서 알아본다. 본 연구의 목표와 관련 연구와의 관계는 그림 1과 같다.

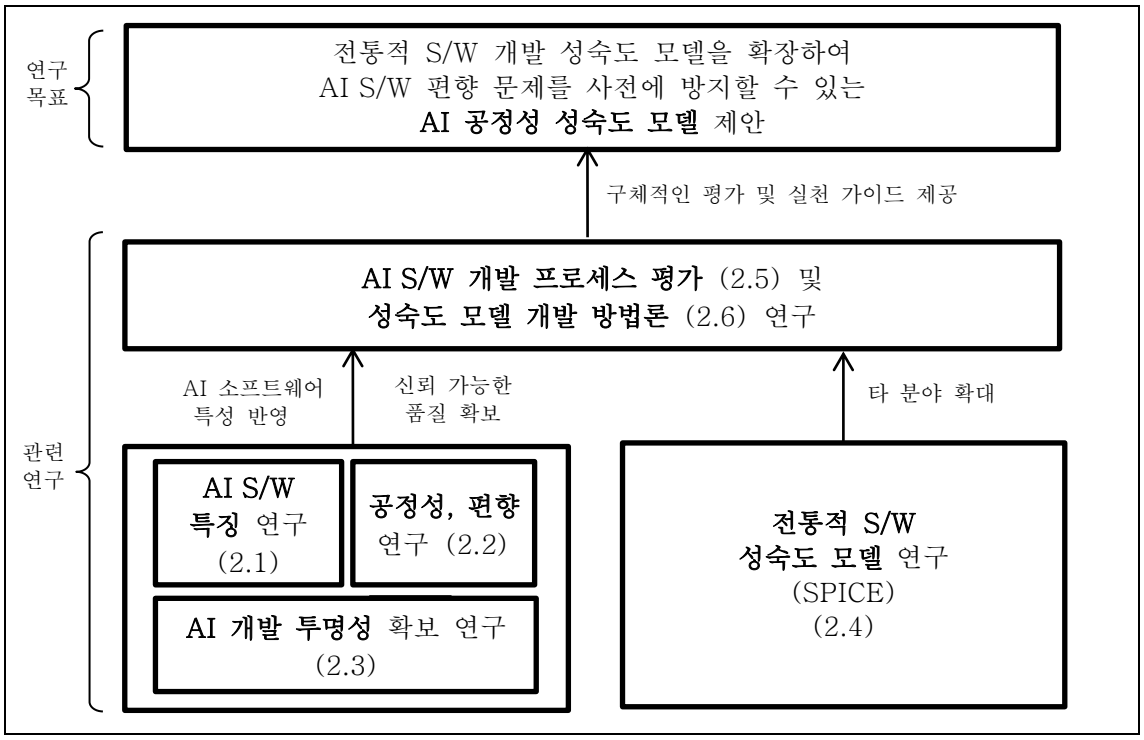

그림 1 연구 목표와 관련 연구

Fig. 1 Research objective and related work

2.1 AI S/W 특징

S/W는 수집된 요구사항을 바탕으로 이를 만족하는 기능을 개발하는 것이 목표이다[3]. 전통적 S/W는 개발자가 직접 요구사항을 분석하고, 이를 기반으로 주어진 입력 데이터에 맞는 출력 데이터와 예외 상황에 대한 대응 방법을 정리한 뒤, 이를 알고리즘으로 공식화하여 기능을 개발한다. 반면 AI S/W는 요구사항을 바탕으로 데이터를 준비하고, AI 모델을 설계한 뒤, 데이터를 학습시키며 알고리즘이 자동으로 공식화된다. 이러한 특성으로 인해 AI S/W는 데이터 중심으로 개발이 진행되고, 도출된 알고리즘은 데이터의 양과 질에 의존성을 갖는다.

Zhang 등[19] 연구에서도 코드 중심의 전통적 S/W 테스트와 달리 데이터 중심의 AI S/W 테스트는 대상, 테스트 결과 확인 방법, 테스트 충분성 등 접근 방법이 다르고, 공정성 등 AI의 특징을 반영한 품질에 대한 연구가 상대적으로 부족하다고 언급하였다.

2.2 공정성 및 편향

공정성이란 AI 가 사람들에게 차별 없는 동작을 제공하는지에 대한 판단 기준으로, AI의 동작 목적에 따라 세부적인 정의는 달라질 수 있다. 사람들은 AI가 항상 공정한 의사결정을 할 것이라고 기대하지만, 실제로는 공정하지 못한 편향된 판단을 하기도 한다.

여기서 편향이란 어떠한 대상이 한쪽으로 치우친 성질을 의미하며, 학습에 활용하는 데이터에도 편향이 존재할 수 있다. 데이터에 존재하는 편향은 목적에 따라 의도적일 수도 있으나, 의도하지 않은 데이터의 편향은 AI S/W의 공정한 동작에 악영향을 미칠 수 있다[20].

편향의 종류는 데이터의 특성에 따라 다양한데, Mehrabi 등[14] 연구에서는 과거 포춘 500 CEO 중 5%만이 여성이었던 것을 고려하지 않아 CEO의 특성에 남성의 특성이 많이 반영된 것처럼 역사적으로 존재하는 데이터의

영향으로 발생하는 "역사적 편향"과 서양 국가를 대상으로 사람의 사진을 수집할 경우 백인의 특징을 더 크게 고려하게 되는 "대표 편향" 등 편향의 유형을 23종으로 분류했고, 이러한 AI 편향을 방지하기 위해서 널리 사용되는 공정성 정의 (기회의 동등성, 가능성의 동등성 등) 10종을 조사했고, 번역, 분류, 예측 등 특정 도메인에서의 편향 완화 기법을 소개하고 있다. 예를 들어, 번역 과정에서 프로그래머는 남성으로, 하우스메이커는 여성으로 번역되는 성별에 따른 편향을 해결하기 위해서, 성 중립적 단어를 활용하거나 남성용 단어와 여성용 단어를 맞바꾼 문장을 추가로 학습하는 등의 방법을 통해 편향을 완화할 수 있다고 소개한다. 이와 같이 도메인별 특정 편향 문제를 해결하는 완화 기법이 일부 존재하지만 아직 초기 연구 단계이고, 특히 조직 차원에서 다양한 범주의 공정성 이슈를 사전에 방지하기 위한 체계에 대한 연구는 부족한 상황이다.

2.3 개발 투명성

AI의 편향을 탐지하고 신속하게 대응하기 위하여 AI 개발에 활용된 데이터 혹은 모델에 대한 투명한 정보를 제공할 필요성이 높아지고 있다. 더욱이, AI 모델은 입력 데이터를 학습하는 과정에서 수억에서 수천억 개가 넘는 내부 변수들이 조정되기 때문에, 그 안의 내용을 개발자가 분석하는 것은 매우 어렵다. AI에 대한 신뢰를 높이기 위해서는 AI가 어떻게 의사결정을 하였는지를 설명할 수 있어야 한다[21].

AI 모델 및 학습에 사용한 데이터의 특성을 투명하게 전달하기 위하여, 구글과 마이크로소프트에서는 학습된 AI 모델의 이름, 개발 목적, 활용 가능한 도메인, 입출력 데이터 양식 등 AI 모델을 사용하기 위한 정보와 성능 평가 결과를 제공하는 AI 모델카드[22]를 제안했고 데이터의 수집 방법, 종류, 사용 목적, 의도된 활용 방법 정보를 제공하는 데이터카드[23]도 제안했다.

2.4 전통적 S/W 성숙도 모델

전통적 S/W의 개발 능력을 평가하고 개선하기 위한 대표적 성숙도 모델에는 SPICE (ISO/IEC 15504)와 CMMI가 있다[24]. 이 중 국제 표준인 SPICE는 S/W 개발 프로세스 수행 능력 심사 및 개선을 지원하고, 특히 개별 프로세스 관점에서 평가와 개선이 용이한 구조를 제공하고 Automotive SPICE[25]와 같이 다른 산업 분야에서도 실효성과 확장성이 검증된 모델이다. SPICE는 그림 2와 같이 프로세스 차원과 능력 차원의 2차원 구조를 제공한다[26]. 프로세스 차원은 조직에서 수행되는 프로세스 단위로 해야 할 활동을 정의한다. 능력 차원은 각 프로세스를 개선하기 위해 필요한 수행 능력을 불완전(0)에서부터 최적화(5)까지 6단계로 구분하여 프로세스의 성숙도 수준을 평가한다.

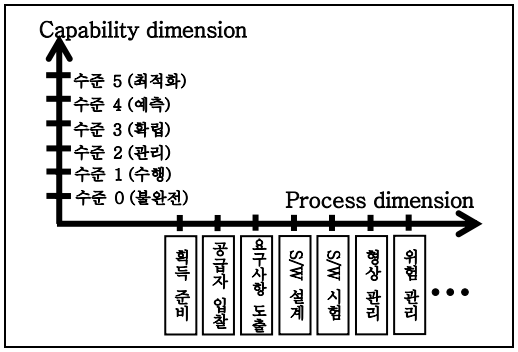

그림 2 SPICE의 2차원 구조

Fig. 2 Two dimensions of SPICE

2.5 AI S/W 개발 프로세스 평가 사례

최근 AI S/W의 개발 프로세스를 평가하기 위한 연구가 활발하게 진행되고 있다.

구글에서는 신뢰성 있는 AI 시스템 개발 능력을 평가하기 위하여, AI S/W의 검증과 관련한 4개 영역 (피처 & 데이터, 모델, 인프라, 상품화 모니터링)을 대상으로 총 28개의 테스트 항목을 도출하고, 이를 기반으로 다음과 같이 상품화 준비 점수(ML Test Score) 산출 방법을 정의하였다[15].

- "테스트 항목" 점수: 미수행 0점, 수동 수행 0.5점, 자동 수행 1.0점

- "영역" 점수: 각 영역에 속하는 테스트 항목의 합산. 각 영역별로 최대 7개 항목으로 최대 7점
- "시스템 최종" 점수: 4개 영역 점수 중 최소 점수

이러한 상품화 준비 점수는 구글의 AI 개발 경험에 기반한 체크리스트 역할을 수행할 수 있고 신뢰성 있는 AI 시스템의 상품화 수준을 파악할 수 있는 정량적 지표로도 활용될 수 있지만, 구체적인 개선 가이드를 제공하지는 않는다.

Amershi 등[16] 연구에서는 마이크로소프트의 AI개발 과제에 적용된 프로세스에 대해 개발자와 관리자에게 수행한 설문조사 결과를 공개하였는데, 여러 과제에서 공통적으로 해결이 필요한 이슈 중 가장 중요한 것으로는 데이터 수집, 전처리, 추적 등 데이터 관리 체계를 구축하는 것이 있었다.

최근 IBM에서도 AI 모델 개발 경험을 반영한 AI 성숙도 프레임워크를 제안하였다[17]. 학습, 검증, 배포, 공정성, 투명성 등으로 정의된 프로세스 차원과 1단계(초기)~5단계(최적화)의 능력 차원으로 구성되어 있다. IBM이 AI 상품화 과제를 개발하면서 습득한 경험을 반영하여 개발 생명주기 전체에 해당하는 AI 프로세스 및 일부 모범 사례를 제공하고 있으나, 공정성과 관련된 내용이 부족하고, 추상적으로 다루고 있어 실용적인 가이드로 활용하기 어렵다.

2.6 성숙도 모델 개발 방법론

Becker 등[18] 연구에서는 IT 분야에서의 빠른 요구사항 변화에 대응하기 위하여 다양한 신규 IT 성숙도 모델이 제안되고 있지만, 기존 모델과 큰 차이점이 없거나 필요한 정보들이 누락된 경우가 많음을 지적하면서, 체계적인 신규 성숙도 모델을 개발하기 위해 지켜야 하는 개발 절차를 제안하였다. 이 과정에서 SPICE, CMMI와 같은 대표적인 모델 6개를 분석하여 성숙도 모델을 개발하는데 필요한 절차를 정의하였다. 주요 개발 절차는 아래와 같다.
- 문제 정의: 성숙도 모델이 풀어야할 도메인과 문제를 정의하고 실수요가 있음을 확인한다.
- 개발 전략 결정: 기존 성숙도 모델과의 비교 분석을 통해 새로운 모델을 개발할지, 아니면 기존 모델을 개선하거나 통합할지 등에 대한 전략을 결정한다.
- 반복적인 성숙도 모델 개발: 기본적인 모델 구조를 설계하고, 모델 세부 속성을 구체화하기 위해 (1) 기존 연구 문헌을 참고하거나 (2) 전문가의 경험적 지식을 활용하여 성숙도 모델을 개발한 뒤, 이해도, 일관성, 문제 적합성에 대한 평가를 받아 보완하는 절차를 반복 수행한다.

3. AI 공정성 성숙도 모델 설계

본 연구에서는 공정하고 신뢰할 수 있는 AI S/W 품질을 확보하기 위해 전통적 S/W 성숙도 모델인 SPICE를 기반으로, 앞서 설명된 Becker 등[18]에서 제시된 성숙도 모델 개발 절차에 따라 최신 연구 동향을 분석하고, 70명의 AI 전문가를 대상으로 설문조사와 인터뷰를 4차례에 걸쳐 수행하여 AI 공정성 성숙도 모델을 정의한다 (그림 3).

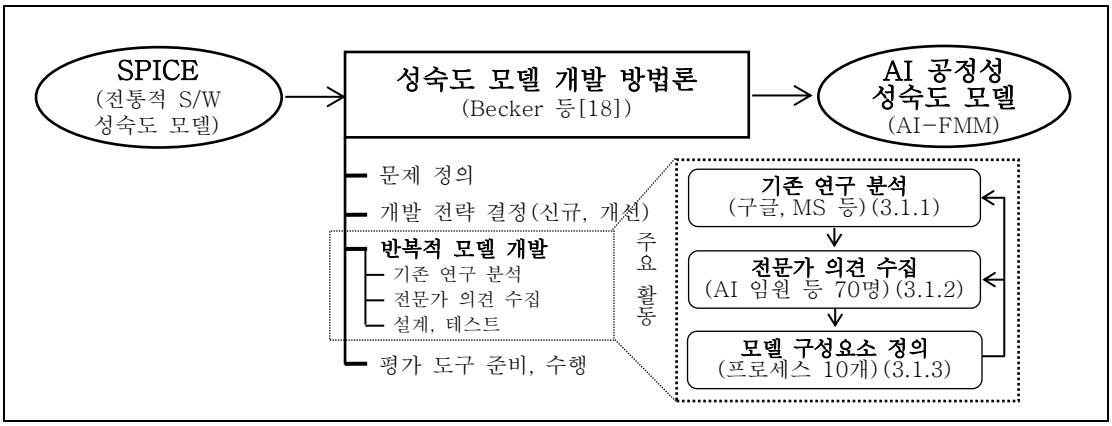

그림 3 AI 공정성 성숙도 모델 설계 과정

Fig. 3 Development approach of AI-FMM

3.1 프로세스 차원 (Process dimension) 설계

그림 3에서 보듯이, Becker 등[18]의 성숙도 모델 개발 절차에서 강조한 주요 활동인 기존 연구 분석, 전문가 의견 수집, 성숙도 모델의 구성요소인 프로세스에 대한 정의를 반복적으로 수행하여 AI 공정성 개발 프로세스를 도출하였다.

3.1.1 기존 연구 분석

공정성 성숙도 모델에 대한 연구가 회사 및 국가 차원에서 논의되고 있지만 아직 초기단계이다. 본 연구에서는 아래와 같이 관련 연구를 분석하여 공정성 프로세스와 세부 기본 사례를 도출하였다.

구글[15]에서는 신뢰성 있는 AI 시스템 개발 능력 평가 체계로 AI S/W 검증에 관련된 4개 영역(피처 및 데이터, 모델, 인프라, 상품화 모니터링)의 28개 항목을 제시하였다. 이 중 서비스 운영 시 발생하는 이슈를 모니터링 하는지, 이슈 발생 시 이전 버전으로 빠르게 복구(rollback)할 수 있는지 등의 내용을 참고하여 "모델 품질 모니터링"의 기본 사례를 도출하였다.

마이크로소프트[27]에서는 AI 개발 단계를 6개 영역(계획, 정의, 프로토타입, 빌드, 출시, 진화)으로 구분하고 관련 48개 세부 항목을 도출한 AI 공정성 체크리스트를 제공하였다. 이 중 "AI 공정성을 확보하기 위해서는, 모델에 포함될 수 있는 잠재적 위협을 도출한 후, 이해관계자들이 검토하고 합의하여 공정성 기준을 정의해야 하며, 평가 가능한 공정성 지표가 준비되어야 한다"는 내용을 참고하여 "공정성 위험요소 분석"과 "공정성 평가 지표"의 기본 사례를 도출하였다.

2022 신뢰할 수 있는 인공지능 개발 안내서 (안)[28]에서는 AI 모델 개발의 5단계 생명주기(계획 및 설계, 데이터 수집 및 처리, AI 모델 개발, 시스템 구현, 운영 및 모니터링)에 따라 총 59개의 점검 항목을 제시하였다. 이 중 데이터의 편향 확인 및 완화 방안, 데이터 축적과 환경 변화에 따른 AI 모델의 성능저하 및 공정성 유지 확인과 같은 내용을 참고하여 "데이터 편향성 평가", "모델 학습 중 편향 검토", "모델의 공정성 유지 검토" 기본 사례에 반영하였다.

그 외에도 데이터 윤리 프레임워크[29]와 데이터 과학자를 위한 윤리 체크리스트[30]를 검토하여 데이터 및 AI 모델 관련 공정성 세부항목을 보완하였다.

3.1.2 전문가 의견 수집

AI와 관련된 업무의 경력이 3년 이상인 전문가 70명을 대상으로 델파이 설문 기법을 이용하여 AI 공정성 관련 의견을 수집했다. 인터뷰 그룹에는 AI 윤리 협의체 및 AI 개발팀에 있는 AI 임원, AI 프로젝트 실무 리더(아키텍트 등), AI 실무 개발자 및 AI 품질 리더 등을 포함하였다(표 1).

표 1 AI 전문가 인터뷰 참석자 정보

Table 1 Stakeholders interviewed for AI processes

업무 영역	임원	실무 리더	실무 개발자	품질 담당자
자연어처리	2	7	4	4
비전인식	–	4	3	
AI 서비스 및 데이터 분석	2	3	7	
뉴럴 심볼릭 및 온디바이스 AI	2	3	3	
로봇, AR, 보안 등	2	6	12	6

1차 설문은 3.1.1절 (기존 연구 분석)을 통해서 도출한 38개의 기본 사례에 대해 설문지와 구두 인터뷰를 통해 전문가 의견을 수집하였다. 수집 결과, AI 개발 및 품질 관련 일반적인 기본 사례가 다소 많고, 일부 중복도 존재

하며, 구체적인 예시나 설명이 부족하다는 의견들이 있었다. 이를 기반으로 기본 사례를 총19개로 조정하였고 예시도 보강하였다.

2차 설문은 AI 공정성 관련 프로젝트의 실무 리더, 실무 개발자, 품질 리더 16명 대상으로 19개 기본 사례에 대하여 필요, 불필요, 기타(판단불가)의 3개 항목으로 의견을 받았다. 전문가의 의견 중 필요보다 불필요 비율이 50% 이상 높은 6개 항목의 경우에는 추가 심화 인터뷰를 통해 평가하여 삭제하거나 통합하였다(표 2).

1, 2차를 포함하여 총 4차례에 걸친 설문조사와 인터뷰를 통해 수집된 의견을 기반으로 기본 사례 도출 및 상세 내용 보완 과정을 반복 수행하였다.

표 2 전문가 설문조사 과정

Table 2 Expert survey process

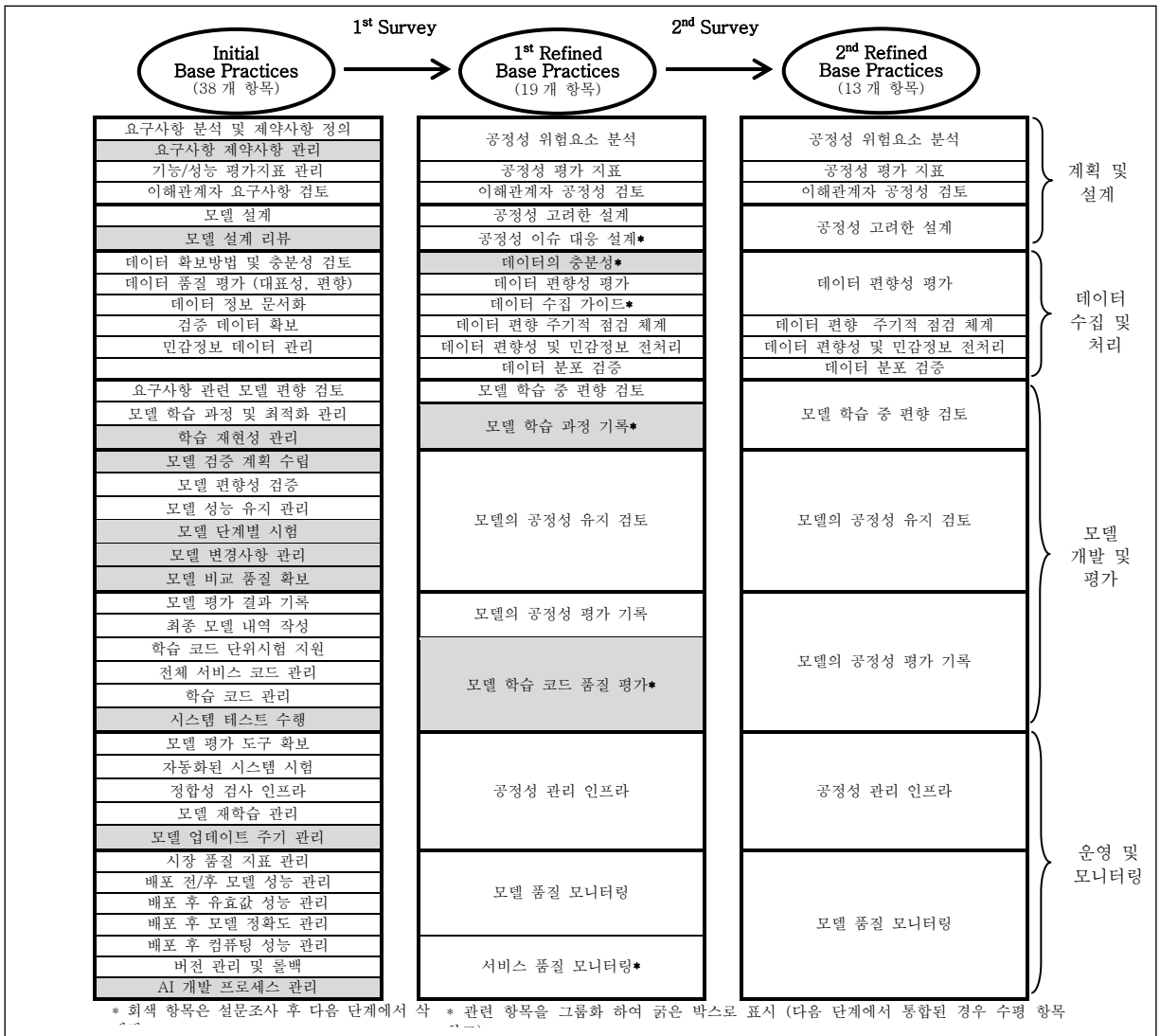

표 3 AI 공정성 프로세스 및 기본 사례 (최종)

Table 3 AI fairness processes and base practices

Majorly changed base practices (*) after 2nd survey	Expert Opinion (불필요 50% 이상 항목 대상)				Final Result
	필요	불필요	기타	Detailed Opinion (including In-depth Interviews)	
공정성 이슈 대응 설계	6	9	1	공정성 사전 방지 설계와 사후 이슈 처리 설계는 함께 고려하는 것이 효율적이어서 "**공정성 고려한 설계**"로 통합 필요함	**통합**
데이터 수집 가이드	5	11		습에 사용되는 데이터도 부족한 경우가 많아 공정 "**데이터 편향성 평가**"와 통합하여 관리 필요함	**삭제** **통합**
모델 학습 과정 기록	6	8	2	AI S/W 개발 특성상 중간 산출물 기록은 의미 없는 경우가 많아 최종 산출물 기록, 관리 위주로 진행 필요	**삭제**
모델 학습 코드 품질 평가	7	7	2	추가 심화 인터뷰 결과, "공정성 항목으로 부적합" 의견이 더 많음	**삭제**
서비스 품질 모니터링	7	7	2	언제든 필요한 AI 모델로 복구할 수 있어야 하고 서비스와 모델로 구분하지 말고 "**모델 품질 모니터링**"에 통합하여 관리 필요	**통합**

(Final)

Process Group	Process	Fairness Base Practice (13개)	
		Name	Description
계획 및 설계	S/W 요구사항 분석	**공정성 위험요소 분석**	AI 모델에서 발생할 수 있는 공정성 관점의 위험요소를 도출 및 상세 범위 설명 작성 ※사례 : "병신"과 같이 사용자가 수치심을 느끼거나 민감하게 반응하는 단어 관리
		공정성 평가 지표	공정성을 평가하기 위한 지표와 평가 기준, 평가 방법을 작성
		이해관계자 공정성 검토	공정성의 기준 및 범위 설정 시, 이해관계자를 명시하고, 정책, 기준 검토에 참여 ※ 이해관계자 : 서비스에 직간접적으로 영향을 받는 사람 (인종, 성별, 연령 등 고려)
	S/W 아키텍처 설계	**공정성 고려한 설계**	공정성 위험요소를 제거 및 방지한 설계, 이슈 발생시 영향 완화를 고려한 설계 (모델 교체 방안 또는 편향 완화를 위한 후처리 기법을 포함한 체계 도입 등)
데이터 수집 및 처리	데이터 수집	**데이터 편향성 평가**	인적, 물리적 요인의 편향을 평가 후 데이터를 수집하고, 필요시 완화 방안을 작성 ※사례 : 편향성 평가가 완료된 국립국어원의 "모두의 말뭉치"를 학습에 사용
	데이터 정제	**데이터 편향 주기적 점검 체계**	데이터의 편향 가능성을 정기적으로 내외부 전문가를 통해 객관적으로 점검 ※사례 : 전문 번역사 그룹을 활용하여 편향을 매 릴리즈마다 점검함
	데이터 전처리	**데이터 편향성 및 민감정보 전처리**	데이터 레이블링 시, 편향 확인 및 전처리(예, 관계없는 데이터 블러 처리) 기법 도입 ※사례 : 데이터 익명화, 관계없는 데이터의 제거, 이미지의 모자이크 처리
		데이터 분포 검증	편향 방지를 위해 데이터 샘플링과 같은 방법으로 데이터 분포 검증 수행
모델 개발 및 평가	학습 과정 관리	**모델 학습 중 편향 검토**	개발하려는 모델에 맞게 편향을 검토하고 제거하는 기법 선택 ※사례 : 다양한 검증 데이터 셋을 준비하여 편향 및 과적합이 발생하였는지 확인
	모델 성능 평가	**모델의 공정성 유지 검토**	시간이 지나도 모델 공정성이 유지되는지 확인하기 위해 모델 변화 모니터링
	최종 선택 AI 모델 관리	**모델의 공정성 평가 기록**	공정성 기준 및 범위, 공정성 평가결과, 이해관계자 그룹별 기대 혜택과 잠재적 손해 간의 상충관계, 모델의 단점, 한계 및 편향을 작성
운영 및 모니터링	AI 인프라	**공정성 관리 인프라**	학습된 모델의 공정성 문제, 학습 데이터의 편향 등 문제 발생시 빠른 업데이트를 위한 지속학습 시스템 확보
	AI 모델 운영 관리	**모델 품질 모니터링**	서비스 후, 예상치 못한 공정성 관련 위험 발생 및 모델의 오사용을 감지 모니터링 하는 경고 기능을 제공하고, 이전으로 되돌리거나 시스템을 긴급 정지할 수 있는 방법 제공

3.1.3 공정성 프로세스 정의

3.1.2절에서 설명한 전문가 의견 수집을 바탕으로, AI S/W 의 공정성 위험을 사전에 빠르게 점검할 수 있도록 최종적으로 4개 프로세스 영역 (① 계획 및 설계 ② 데이터 수집 및 처리, ③ 모델 개발 및 평가, ④ 운영 및 모니터링)의 10개의 프로세스를 대상으로 총 13개의 공정성 기본 사례를 표 3과 같이 정의하였다.

3.2 능력 차원 (Capability dimension) 설계

SPICE의 경우 표 4에서 보듯이, 능력 수준 불완전(0)부터 최적화(5)까지 6단계의 능력 차원 정의를 제공하고 있고 이를 기반으로 각 프로세스의 능력 수준은 기본 사례 평가를 통해서 결정된다[31].

표 4 능력 차원

Table 4 Capability dimension

Capability Level		Description
수준 5	최적화 (Optimizing)	프로세스 수행이 지속적으로 최적화되고 개선된다
수준 4	예측 (Predictable)	정량적으로 수행을 예측하고 관리한다
수준 3	확립 (Established)	조직 표준으로 정의된 프로세스를 사용한다
수준 2	관리 (Managed)	관리된 방식으로 수행, 산출물 수립, 통제, 유지된다
수준 1	수행 (Performed)	수행된 프로세스가 프로세스 목적을 달성한다
수준 0	불완전 (Incomplete)	프로세스가 수행되지 않거나 목적을 달성하지 못한다

본 연구에서는 구체적인 AI 개발 활동에 대한 확장이 필요했던 프로세스 차원과는 달리, 능력 차원의 경우 SPICE의 정의를 동일하게 준수하고자 한다. 이는 SPICE의 대표적인 확장 사례인 Automotive SPICE[25]의 경우에도 동일하고, 추후 공정성 뿐 아니라 안전성, 강건성과 같이 AI S/W가 만족해야 하는 다양한 품질 속성에 대한 확장 용이성을 고려하였다.

예를 들어, 안전성(safety) 성숙도 모델을 정의하기 위해서는 공정성 성숙도 모델을 정의한 것처럼 Becker 등[18]의 개발 방법론에 따라 기존 연구 분석, 안전성 전문가 인터뷰, 안전성 기본 사례 정의와 같은 절차를 통해서 AI 안전성 프로세스를 확장 정의할 수 있다. 대표적으로, "데이터 전처리" 프로세스 대상으로 AI S/W가 외부의 악의적인 공격으로부터 적절하게 대응가능한지를 평가하기 위하여 "데이터 공격 점검" 기본 사례를 정의할 수 있다. 악의적인 공격의 예시로는 CCTV의 번호판 인식에 필요한 이미지의 일부분을 인위적으로 변조하여, 사람이 보기에는 정상적으로 인식하지만 AI S/W는 이를 제대로 인식하지 못하게 하는 것이 있다. 이러한 공격에 대응하기 위

해서는 학습 및 검증 데이터에 대한 공격을 방어하는 대책을 준비할 필요가 있고, 이를 안전성 기본 사례로 정의할 수 있다.

4. 사례 연구를 통한 실효성 분석

본 연구에서 정의한 AI 공정성 성숙도 모델을 현재 개발 및 운영 중인 과제에 적용함으로써 공정성 수준에 대한 평가 능력이 있는지, 각 과제의 특성을 반영한 구체적인 개선 가이드를 제시할 수 있는지, 그리고 실제 공정성 이슈의 감소에 기여하는지를 분석하였다.

4.1 공정성 평가 능력

본 성숙도 모델의 공정성 평가 능력을 확인하기 위해 사례 A에 적용된 평가 결과를 분석한 결과, 프로세스별 공정성 수행 활동에 맞게 수준을 평가하는 것으로 확인할 수 있었다.

사례 A는 자연어 처리 기반의 AI 서비스로써 평가 결과는 표 5에서 보듯이, 10개 프로세스 중 공정성 능력 수준 3으로 평가된 것은 4개, 수준 2는 3개, 수준 1은 1개, 수준 0은 2개이다. 프로세스 평가 결과의 대표적인 예는 다음과 같다.

표 5 AI 공정성 성숙도 모델 기반 평가 결과 (사례 A)

Table 5 An assessment result of a case study A using AI-FMM

Process dimension			Capability dimension						Assessment details
Process Group	Process	Base Practice	수준 1	수준 2	수준 3	수준 4	수준 5	Result (Process)	
계획 및 설계	S/W 요구사항 분석	공정성 위험요소 분석	O					수준 1	욕설, 젠더 편향과 같이 공정성 기준을 제공
		공정성 평가 지표	O						평가셋에 공정성 평가 테스트 케이스를 포함하고 있음
		이해관계자 공정성 검토	O						4 개 국어는 전문 번역사의 자문을 받고 있음
	S/W 아키텍처 설계	공정성 고려한 설계	O	O	O			수준 3	공정성 이슈를 위한 전처리, 후처리 필터링 설계됨
데이터 수집 및 처리	데이터 수집	데이터 편향성 평가	O	O	O			수준 3	편향성 평가를 거친 국립국어원의 "모두의 말뭉치" 활용
	데이터 정제	데이터 편향 주기적 점검체계	O	O	O			수준 3	릴리즈 전 점검 및 데이터 편향 제보 있을 경우 부서내 검토 진행
	데이터 전처리	데이터 편향성 및 민감정보 전처리	O	O	O			수준 3	이메일이나 전화번호 등 사용자 민감정보는 문장을 필터링하여 삭제, 관리
		데이터 분포 검증	N/A						AI 번역 특성상 데이터의 분포 검증 불필요
모델 개발 및 평가	학습 과정 관리	모델 학습 중 편향 검토	O	O				수준 2	다양한 학습 데이터 활용하여 모델의 과적합과 같은 편향 완화 작업 수행
	모델 성능 평가	모델의 공정성 유지 검토						수준 0	모델 공정성 유지 평가 미수행
	최종 선택 AI 모델 관리	모델의 공정성 평가 기록						수준 0	공정성 평가 기록 미수행
운영 및 모니터링	AI 인프라	공정성 관리 인프라	O	O				수준 2	공정성 관리를 위한 지속적 재학습 (Continuous Learning) 체계 확보됨
	AI 모델 운영 관리	모델 품질 모니터링	O	O				수준 2	실사용 로그 기반 AI 모델 품질 모니터링 수행

- "S/W 요구사항 분석" 프로세스의 경우 공정성 기준과 평가 테스트 케이스가 있고 전문 번역사의 자문을 받고 있어 수준 1의 기준을 만족하나, 공정성 기준에 대한 변경 관리가 없고, 요구사항과 연계된 테스트 케이스의 이력 관리가 이루어지지 않아 수준 2로는 평가되지 못했다.
- "학습 과정 관리" 프로세스의 경우, 학습 데이터 중 편향이 우려되는 출처는 배제하였고, 모델카드와 같은 산출물로도 이력 관리되어 수준 2의 기준을 달성하였다. 그러나 조직내 다른 사례에서는 일관성 있게 수행되지 않아 수준 3으로 평가되지 않았다. "AI 인프라" 프로세스의 경우에도 운영 시스템은 확보되었으나, 표준 운영 가이드가 부족하여 수준 2로 평가되었다.
- "데이터 전처리" 프로세스의 경우, 자연어 처리 도메인의 특성상 공정성 위험에 대한 사전 대비의 필요성이 부각되어, 수준 1에서 수준 3까지에 해당하는 활동을 사전에 수행하고 있었다. 즉, 데이터를 전처리하는 과정에서 편향을 제거하고 욕설을 필터링하고 있었으며, 모델카드와 같은 산출물로 처리 내역을 관리하고, 조직내 AI 개발 프로세스도 확립되어 있어 수준 3의 기준을 만족하나, 데이터 전처리 결과를 정량적으로 모니터링하고 지속

적으로 개선하는 체계를 구축하지 못해 수준 4로는 평가되지 못하였다.

참고로, 사례 A의 경우 AI 공정성 활동 및 산출물을 투명하게 제공하는AI 모델카드(표 6)를 제공함으로써, 5개 프로세스가 능력 수준 2 이상으로 평가받는데 실질적으로 기여함을 확인할 수 있었다. 예를 들어, "S/W아키텍처 설계" 프로세스는 "번역 결과 출력 시 사용자에게 불쾌감을 줄 수 있는 욕설에 대해 필터링 하는 기능은 탑재하고 있음"과 같은 공정성 이슈에 대한 대응 정책과 설계 정보를 AI 모델카드를 통해 제공하여 수준 2 이상으로 평가되었다.

표 6 AI 모델카드 일부 (사례 A)

Table 6 A part of AI model card (Case study A)

Item		Description
모델	목적	클라우드 번역 서비스를 제공하기 위해 개발된 텍스트 번역 모델로 8개 언어 (Ko,En 등) 대상 서비스 제공함
	아키텍처	AI 번역을 위한 트랜스포머(Transformer) 기반 모델
	입출력	- 입력: 소스 언어, 목적 언어, 입력 문장 (UTF-8 형식) - 출력: 목적 언어 번역 결과 텍스트 (UTF-8), 처리 상태
학습 데이터	특징	- 모두의 말뭉치(국립국어원), 구매 데이터 등 - 데이터 분량 1억 문장 이상
	학습 방법	중복 제거 및 노이즈 제거 필터링, 하위 단어 토큰화 등 전처리,엔비디아 V100 GPU 32개 활용, Dropout 0.05
테스트 데이터	특징	F사 3000 문장 구매
	테스트 방법	정의된 번역 API를 통해 AI 모델 접근, 번역 결과 추출 후 스크립트 통해 BLEU score(SacreBLEU도구 활용) 측정
성능 지표	이름 및 계산식	BLEU(BiLingual Evaluation Understudy) score: 번역 품질 지표. 계산식:https://en.wikipedia.org/wiki/BLEU
	측정 방법	벤치마크 평가 도구를 활용하여 실행 및 분석 (API 호출로 번역 후 정답문 비교 측정)
	결과	BLEU score: 주요 기업 대비 동등, 우위 수준
윤리 (공정성)		- 주어, 목적어 등이 누락된 입력문에서 목적 언어에 따라 임의의 대명사로 복원되는 과정에서 편향성 발생 가능 - 유명 정치인이나 셀럽 등의 경우 번역시 의도하지 않은 별명, 별칭으로 번역됨에 따라 정치 편향성 발생 가능 - 학습 데이터 출처별 성향을 고려해 편향성이 우려되는 출처는 배제 - AI 번역은 주어진 입력에 대해 목적 언어로 충실하게 변환하는 태스크로 입력에 대한 판단하지 않음 - AI 번역의 결과 출력 시 사용자에게 불쾌감을 줄 수 있는 욕설에 대해 필터링하는 기능은 탑재하고 있음

4.2 개선 가이드 제공

본 성숙도 모델이 AI 공정성 문제 해결을 위해 필요한 개선 가이드를 실효적으로 제시할 수 있는가를 확인하기 위해, 개인의 생활 데이터를 기반으로 한 추천 서비스를 운영 중인 과제 대상으로 사례 연구를 수행하였다. 해당 과제는 일부 프로세스 대상으로 낮은 수준의 평가를 받았는데, 해당 항목들에 대하여 표 7과 같이 구체적인 개선 가이드를 제공하였다.

표 7 공정성 개선 가이드 사례 (사례 B)

Table 7 Fairness improvement guides (Case study B)

Process Group	Process	Base Practice	Improvement Guide
계획 및 설계	S/W 요구사항 분석	공정성 위험요소 분석	공정성 위험요소를 분석하고, 공정성 범위 선정을 통해 민감어를 지정하여 예방이 가능
		이해관계자 공정성 검토	인종, 국가, 성별 등을 고려하여 다양한 이해관계자를 명시한 후, 요구사항 수집 시 참여시킴
데이터 수집 및 처리	데이터 수집	데이터 편향성 평가	데이터 수집 시 데이터에 적절한 레이블링 및 전처리 되었는지, 또 사용자 패턴에 맞춰 균등하게 수집되었는지 확인함
운영 및 모니터링	AI 인프라	공정성 관리 인프라	공정성 관련 위험 발생 모니터링을 하기 위하여, 시장 VOC에 대한 피드백 체계 운영 등을 통해 이슈 발생시 즉각 대응 및 지속적 재학습 및 배포 수행

개발자 인터뷰를 통해서 아래와 같이 제공된 공정성 개선 가이드가 공정성 이슈를 개선하는데 실질적으로 도움이 되었으며, 해당 가이드를 후속 AI 모델 개발에도 반영하고 있다는 피드백을 받았다.

- 이해관계자 공정성 검토: 다양한 이해관계자와 사전 협의 없이 개발자 자의로 공정성 기준을 수립하여, 특정 국가의 관습을 고려하지 않은 차별 이슈가 발생하였다. 이를 개선하기 위해 요구사항을 수집할 때 인종, 성별, 연령, 지역, 언어, 종교, 장애 등을 고려한 다양한 이해관계자의 의견을 청취하고 정책에 반영하도록 가이드 하였다.
- 데이터 편향성 평가: 서비스 운영 조직과 AI 모델 개발 조직 사이의 역할이 명확히 정의되지 않아서 데이터 편향성 평가와 데이터 분포 검증을 고려하지 못했다. 예를 들어, 사용자의 생활패턴을 아침, 점심, 저녁으로 구분해야 하는데, 데이터 수집 시 이를 고려하지 못해 편향된 추천을 하는 문제점이 발생하였다. 이를 개선하기 위해, 데이터를 수집할 때, 시간적, 인적, 물리적 요인으로 인한 편향이 있었는지 확인하여 조치하고 데이터의 레이블링, 전처리, 분포 검증을 통해 편향의 가능성을 최소화 할 수 있도록 가이드 하였다.

- 공정성 관리 인프라: AI 모델의 지속학습 시스템은 확보되었으나, 사용자의 데이터를 수집하는 운영 조직과 학습을 수행하는 개발 조직 사이에 역할 구분 및 커뮤니케이션 관련 이슈가 발생하였다. 이로 인해 서비스 조직에서 수집된 사용자의 피드백이 개발 조직에 즉각 공유되지 못하였다. 이를 개선하기 위해, 공정성 이슈 발생을 모니터링하고 이슈가 발생했을 때 재학습 후 배포로 연계하여 즉각 대응할 수 있는 체계를 구축하도록 가이드 하였다.

4.3 공정성 이슈 감소 기여

본 성숙도 모델이 공정성 이슈 감소에 어떤 영향을 미치는지 점검하기 위해 자연어 처리 기반의 AI 서비스 다수를 대상으로 사례 분석을 수행하였다.

공정성 이슈로는 특정 언어의 인식률이 낮거나 외국인을 고려하는 부분이 미흡하는 등 국가, 연령, 성별, 특정 집단 관점의 차별이 존재하는데, 이를 분석해본 결과, 2019년 이후 서비스들의 규모가 지속적으로 커짐에도 불구하고 그림 4에서 보듯이 AI 서비스 처리 건수 백만 개당 검출된 공정성 이슈는 오히려 큰 폭으로 감소하는 것을 확인할 수 있었다.

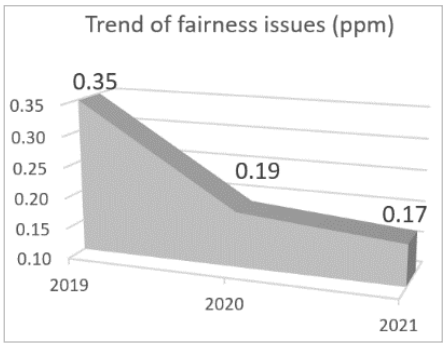

그림 4 연도별 공정성 이슈 추이

Fig. 4 Trend of fairness issues

이러한 2019년 이후 감소 추세는 AI 윤리 원칙, AI 윤리 가이드라인, 민감어 대응 체계, AI 윤리 교육, AI 윤리 협의체 등 공정성 개선 활동[32]의 결과로 보인다. 공정성 개선 활동들은 본 성숙도 모델 개발과 병행으로 진행이 되어 상호보완적인 관계를 가지고 있고, 이는 제안한 성숙도 모델이 실제 공정성 이슈 저감에 기여가 가능함을 보여준다. AI 윤리 원칙 및 가이드라인, 민감어 대응 체계에 대한 설명은 아래와 같다.

- AI 윤리 원칙 및 가이드라인: AI가 인권에 미칠 수 있는 영향 등 윤리적 측면을 고려하기 위해 "공정성", "투명성", "책임성" 이라는 인간에게 이로운 AI를 위한 AI 윤리 핵심 원칙을 2019년 소비자 가전 전시회(CES)에서 공개했다[33]. 또한, AI 윤리 가이드라인으로 "학습 데이터 입력 및 알고리즘 설계 시, 편향성을 제거할 수 있는 전략 및 절차 검토", "학습 데이터 내역, 알고리즘 설계 방법 등 AI 시스템의 생성 과정을 기록" 등 총 23개 항목을 회사내 가이드로 제공하고, 본 성숙도 모델의 8개 프로세스에 해당 내용이 반영되어 있다.
- 민감어 대응 체계: 불공정한 편견이 조장되지 않도록 각 국가의 법규와 사회적 윤리, 소비자 정서 등을 고려하여 민감어 처리 정책을 확립하였다. 종교, 국가, 인종 등 민감어 데이터베이스(DB)를 운영함으로써 AI 공정성 기본 사례의 "공정성 위험요소 분석" 및 "공정성 평가 지표" 관련 활동을 하고 있다. 또한, 지속적 이슈 센싱 및 민감어 탐색 시험을 통해 "모델 품질 모니터링" 및 "데이터 편향성 평가" 관련 활동을 하고 있으며, 관련 내용이 본 성숙도 모델의 6개 프로세스에 반영되어 있다(그림 5).

위와 같은 공정성 개선 활동 결과, 글로벌 지속가능경영 연합체인 World Benchmarking Alliance의 2021년 "디지털 포용성 평가"에서 AI 윤리 원칙 공개에 대해 높은 점수를 받아 글로벌 순위 4위로 평가[34]를 받는 등 대외적인 개선 성과 또한 확인된다.

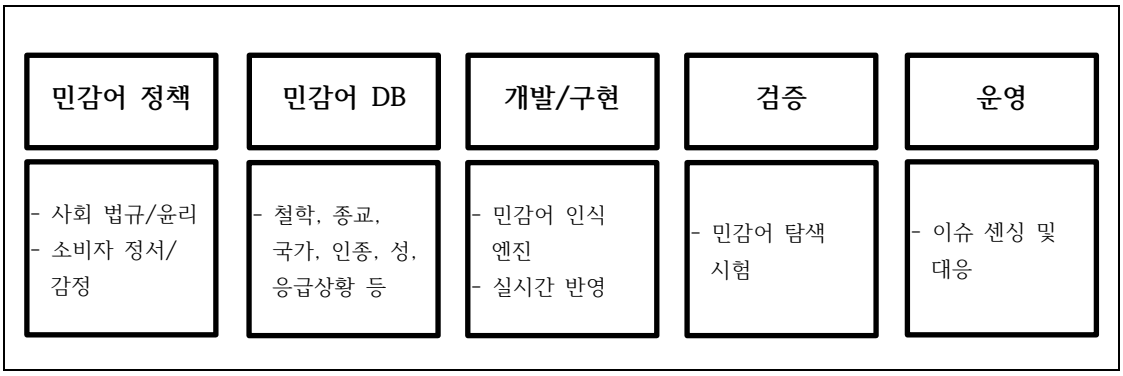

그림 5 민감어 대응 체계

Fig. 5 AI ethics protection system

5. 토론: 공정성 판단 기준

다수의 개발자는 인터뷰 과정에서 시장 VOC 중 공정성 이슈인지, 품질 이슈인지 판단하기 어려운 경우가 많음을 언급하였다. 예를 들면, 인도식 영어와 같은 특정 언어 혹은 지역 방언을 지원하지 않는 것이 편향 혹은 차별에 해당하는 것인지, 아니면 서비스 확장의 제한으로 봐야하는 것인지에 대해서 논란의 여지가 있으니, 구체적인 공정성 판단 기준이 필요하다는 의견을 주었다. 본 연구에서는 하기와 같이 키워드 검색 및 심각도 수준 분류를 통한 공정성 판단 기준을 제안한다.

키워드 검색을 위해, 1장에서 소개한 챗봇 인종 차별, 애플카드 신용한도 차별 등 널리 알려진 공정성 이슈 사례를 참고하여 여성, 노인, 아동, 흑인 등 편향과 관련된 키워드144개를 선정하였고, 이를 활용하여 시장 VOC 중 43개의 공정성 이슈를 파악할 수 있었다. 이후 1차로 검출된 43개 공정성 이슈를 상세히 검토하는 과정에서, 파생되는 공정성 키워드를 추가로 발견하거나 영어 단어 키워드를 보충하는 등의 작업을 수행하였고, 이를 통해 좀 더 체계적으로 키워드 집합을 확장할 필요가 있다고 판단하였다. 이에, 공정한 고용 기회 관련된 구글[35] 및 미연방[36] 문헌을 분석하여 성별, 연령, 국가, 장애, 정치, 종교, 신분, 기타(범죄 등)의 8개 영역을 정의하였고, 각 영역별로 인터넷 검색과 문헌 조사를 통해 공정성 키워드를 추가로 발굴하고, 더불어 사전을 통해 연관 키워드 및 한글과 대응되는 영어 단어를 보강하여 최종 630 개 키워드 집합을 확보하였다. 이렇게 검색 키워드 집합을 144개에서 630개로 4배 가량 늘린 결과, 관련된 공정성 이슈는 43개에서 263개로 6배 확대됨을 확인할 수 있었다. 즉, 검색 키워드 집합을 어떻게 정의하는지에 따라 공정성 이슈 대상이 크게 달라질 수 있음을 확인하였다.

키워드를 활용하여 검출한 공정성 이슈는 사람마다 다르게 평가할 수 있기 때문에, 공정성 이슈를 심각도 수준에 따라 분류하고 우선순위화하여 세밀한 공정성 판단 기준을 정의하는 방안을 정립하게 되었다. 즉, 표 8과 같이 정의된 4개의 심각도 수준을 활용하여 앞에서 키워드 검색으로 검출한 총 263개 공정성 이슈를 심각도 수준 A 26개, B 105개, C 109개, D 23개로 분류하였다.

표 8 공정성 심각도 수준 및 사례

Table 8 Fairness severity levels and examples

Severity Level	Definition	Examples
A	명백한 공정성 이슈이며, 즉시 수정이 필요함	- 특정 언어 억양 인식 정확도 낮음
B	당장의 문제는 아닐 수 있으나 잠재적 리스크 큼	- 중국에서 중국어로 특정 앱을 열어 달라고 했으나 영어로 안된다고 답함
C	상황에 따라 공정성 이슈의 여부가 불분명하나 리스크 존재	- 일기예보와 함께 항상 어두운 뉴스가 나옴 (대량학살, 무거운 정치 이슈)
D	공정성 이슈는 아니나 수정을 권고 함	- 뉴스 제공해주는 서비스 사용시, 뉴스 공급자 변경 옵션이 있었으면

이렇게 분류된 이슈 중 공정성 이슈를 심각도 수준 A만으로 한정할지 아니면 A, B, C, D 를 모두 포함할 것인지에 따라 공정성 이슈의 수는 10배 차이가 날 수 있다. 또한 공정성 판단 기준을 심각도 수준 A로 제한하면 심각

하고 즉시 처리해야 하는 이슈 위주로 사전 리스크를 예방 관리할 수 있고, 심각도 수준 A, B, C, D를 모두 포함하도록 확대하면 제품과 서비스에는 드러나지 않지만, 잠재된 위험성이 내포된 공정성 이슈까지 예방 및 관리가 가능하다.

6. 결론 및 향후 연구

본 연구는 공정하고 신뢰할 수 있는 AI S/W 품질을 확보하기 위한 AI 공정성 성숙도 모델인 AI-FMM을 제안하고, 다수의 AI 서비스를 대상으로 사례 연구를 진행하여 AI-FMM의 실효성을 보였다. 또한, AI 개발 프로세스별 기본 사례 및 구체적 실행 가이드를 통하여 AI S/W의 일관적인 품질 확보가 가능함을 보였다.

제안된 성숙도 모델은 공정한 AI 서비스를 지향하여 기업의 브랜드 가치 및 소비자 신뢰를 제고하는데 기반을 제공한다. 이외에도, AI 개발 능력 검증과 AI 표준 인증 사전 대비를 도와 기업 경쟁력 향상에 도움이 될 것으로 기대된다.

향후 연구로는 AI 공정성 성숙도 모델 적용 확산을 위한 추가 적용 사례 발굴을 포함한다. 번역, 음성 뿐 아니라 비전 등 다양한 AI 도메인에서의 성숙도 모델 적용을 통해 공정성 프로세스 및 상세 기본 사례가 충분한지를 검증하고, 개선하고 보완할 사항이 있는지 점검할 것이다. 또한, 공정성과 유사하게 안전성 등 다른 측면의 품질 속성 영역으로 AI 성숙도 모델 구조를 확장하기 위해, AI 공통적인 프로세스와 공정성, 안전성과 같은 품질 영역에 특화된 가변 프로세스를 고려한 통합 성숙도 모델을 개발하고자 한다. 이러한 연구 방향은 복합적 측면의 품질 보장을 고려해야 하는 서비스 로봇과 같은 AI 제품의 신뢰성 향상에 도움을 줄 것이다.

참고문헌

[1] P. Stone, R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus, K. Leyton-Brown, D. Parkes, W. Press, A. Saxenian, J. Shah, M. Tambe, and A. Teller, Artificial Intelligence and Life in 2030 - One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, pp. 52, Stanford University, CA, 2016.

[2] M. Needhan. (2021, Feb. 23). IDC Forecasts Improved Growth for Global AI Market in 2021 [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=prUS47482321 (downloaded 2022, Sep. 20)

[3] M. Dilhara, A. Ketkar, and D. Dig, "Understanding Software-2.0: A Study of Machine Learning Library Usage and Evolution," ACM Transactions on Software Engineering and Methodology, Vol. 30, No. 4, pp. 1-42, Jul. 2021.

[4] B. Babic, I. Cohen, T. Evgeniou, and S. Gerke, "When Machine Learning Goes Off the Rails", Harvard Business Review, pp.21-32, Jan. 2021.

[5] O. Schwartz. (2019, Nov. 25). In 2016 Microsoft's racist chatbot revealed the dangers of online conversation [Online]. Available: https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation (downloaded 2022, Sep. 20)

[6] J. Dastin. (2018, Oct. 11). Amazon scraps secret AI recruiting tool that showed bias against women [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (downloaded 2022, Sep. 20)

[7] T. Telford. (2019, Nov. 11). Apple Card algorithm sparks gender bias allegations against Goldman Sachs [Online]. Available: https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/ (downloaded 2022, Sep. 20)

[8] N. Kayser-Bril. (2020, Apr. 7). Google apologizes after its Vision AI produced racist results [Online]. Available: https://algorithmwatch.org/en/google-vision-racism/ (downloaded 2022, Sep. 20)

[9] Google. (2018). Artificial Intelligence at Google: Our Principles [Online]. Available: https://ai.google/principles/ (downloaded 2022, Sep. 20)

[10] Microsoft. (2018). Microsoft responsible AI principles [Online]. Available: https://www.microsoft.com/en-us/ai/our-approach?activetab=pivot1%3aprimaryr5 (downloaded 2022, Sep. 20)

[11] PAI. (2016). Partnership on AI [Online]. Available: https://partnershiponai.org/ (downloaded 2022, Sep. 20)

[12] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, pp. 108, European Commission, Brussels, 2021.

[13] Korean Standards Association. (2020). AI+ Certification [Online]. Available: https://www.ksa.or.kr/ksa_kr/6962/subview.do (downloaded 2022, Sep. 20) (in Korean)

[14] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," ACM Computing Surveys, Vol. 54, No. 6, pp. 1-35, Jul. 2022.

[15] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley, "The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction," Proc. of the 2017 IEEE International Conference on Big Data (Big Data), pp. 1123-1132, 2017.

[16] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, "Software Engineering for Machine Learning: A Case Study," Proc. of the IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice, pp. 291-300, 2019.

[17] R. Akkiraju, V. Sinha, A. Xu, J. Mahmud, P. Gundecha, Z. Liu, X. Liu, and J. Schumacher, "Characterizing

Machine Learning Processes: A Maturity Framework," Proc. of the International Conference on Business Process Management, pp. 17-31, 2020.
[18] J. Becker, R. Knackstedt, and J. Pöppelbuß, "Developing Maturity Models for IT Management," Business & Information Systems Engineering, Vol. 1, No. 3, pp. 213-222, May 2009.
[19] J. Zhang, M. Harman, L. Ma, and Y. Liu, "Machine Learning Testing: Survey, Landscapes and Horizons," IEEE Transactions on Software Engineering, Vol. 48, No. 1, pp. 1-36, Jan. 2022.
[20] G. Satell and Y. Abdel-Magied. "AI Fairness Isn't Just an Ethical Issue," Harvard Business Review, Oct. 2020.
[21] IBM Research. (2018). Trusted AI. 2018 [Online]. Available: https://research.ibm.com/teams/trusted-ai (downloaded 2022, Sep. 20).
[22] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. Raji, and T. Gebru, "Model cards for model reporting," Proc. of the ACM Conference on Fairness, Accountability, and Transparency, pp. 220–229, 2019.
[23] T. Gebru, J. Morgenstern, B. Vecchione, J. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for Datasets," Communications of the ACM, Vol. 64, No. 12, pp. 86-92, Dec. 2021.
[24] N. Ehsan, A. Perwaiz, J. Arif, E. Mirza, and A. Ishaque, "CMMI/SPICE based process improvement," Proc. of the IEEE International Conference on Management of Innovation & Technology, pp. 859-862, 2010.
[25] Automotive SIG. (2017). Automotive SPICE Process Reference and Assessment Model 3.1 [Online]. Available: http://www.automotivespice.com/fileadmin/software-download/AutomotiveSPICE_PAM_31.pdf (downloaded 2022, Sep. 20)
[26] K. Emam and H. Jung, "An empirical evaluation of the ISO/IEC 15504 assessment model," Journal of Systems and Software, Vol. 59, No. 1, pp. 23-41, Oct. 2001.
[27] M. Madaio, L. Stark, J. Vaughan, and H. Wallach, "Co-designing checklists to understand organizational challenges and opportunities around fairness in AI," Proc. of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-14, 2020.
[28] Ministry of Science and ICT, TTA. (2022). 2022 Trustworthy Artificial Intelligence Development Guide (draft) [Online]. Available: https://www.nipa.kr/main/downloadBsnsDtlsIemNttFile.do?bsnsDtlsIemNttAtchmnflNo=6864 (downloaded 2022, Sep. 20) (in Korean)
[29] Gov.UK. (2020). Data Ethics Framework [Online]. Available: https://www.gov.uk/government/publications/data-ethics-framework (downloaded 2022, Sep. 20)
[30] DrivenData. (2020). Data Science Ethics Checklist [Online]. Available: https://deon.drivendata.org/ (downloaded 2022, Sep. 20)
[31] Y. Kwon, Y. Ko, J. Kim, and Y. Koo, "Design and Implementation of Assessment System for SPICE Maintenance Process," Journal of KIISE:Computing Practices and Letters, Vol. 8, No. 2, pp. 141-154, 2002. (in Korean)
[32] Samsung. (2022). A JOURNEY TOWARDS A SUSTAINABLE FUTURE – Samsung Electronics Sustainability Report 2022 [Online]. Available: https://images.samsung.com/is/content/samsung/assets/uk/sustainability/overview/Samsung_Electronics_Sustainability_Report_2022.pdf (downloaded 2022, Sep. 20)
[33] Samsung. (2019). Samsung AI principles [Online]. Available: https://www.samsung.com/latin_en/sustainability/digital-responsibility/ai-ethics/ (downloaded 2022, Sep. 20)
[34] J. Lee. (2021, Dec. 8). Samsung ranks 4th on global digital inclusion index [Online]. Available: https://www.koreaherald.com/view.php?ud=20211208000671 (downloaded 2022, Sep. 20)
[35] Google. (n.d.). Policy on harassment, discrimination, retaliation, standards of conduct, and workplace

concerns (US) [Online]. Available: https://services.google.com/fh/files/blogs/policy_workplace_concerns.pdf (downloaded 2022, Sep. 20)
[36] U.S. Department of Labor. (2009). Equal Employment Opportunity Posters [Online]. Available: https://www.dol.gov/agencies/ofccp/posters (downloaded 2022, Sep. 20)