

지능정보사회 이용자 보호 정책개발

2024. 12

연구기관 : 정보통신정책연구원



정책연구 24-17

지능정보사회 이용자 보호 정책개발

조성은 외

2024. 12

연구기관 : 정보통신정책연구원



방송통신위원회

Korea Communications Commission

제 출 문

방송통신위원회 위원장 귀하

본 보고서를 『지능정보사회 이용자 보호 정책개발』의 연구결과보고서로 제출합니다.

2024년 12월

연구기관: 정보통신정책연구원

총괄책임자: 조성은 연구위원

참여연구원: 심홍진 연구위원

문정욱 실장

이시직 부연구위원

김지혜 전문연구원

김희연 부연구위원

박지원 연구위원

이지호 연구위원

연제선 연구위원

이훈 교수

정낙원 교수

이대호 교수

선지원 교수

정창우 변호사

목 차

요약문	xiii
제1장 서론	1
제1절 연구의 필요성 및 목적	1
1. 방송통신위원회의 기본 원칙	1
2. 가이드라인 개발의 필요: 생성형 인공지능 서비스의 확산과 이용자 권익 침해 이슈	4
3. 자율규범의 충실한 이행: 사업자의 기본원칙 준수 상황 점검	6
제2절 연구 주요 구성	7
제2장 생성형 인공지능 서비스 규제 동향과 이용자 연구	9
제1절 생성형 인공지능 서비스 가이드라인 동향	9
1. 공공기관의 생성형 인공지능 활용 가이드라인	9
2. 대학의 생성형 인공지능 활용 가이드라인	15
제2절 생성형 인공지능 서비스 규제 동향	22
1. 국제기구	22
2. 유럽연합(EU)	28
3. 미국	33
4. 영국	45
5. 캐나다	50
6. 일본	53
7. 우리나라	57
제3절 소 결	61
제3장 생성형 인공지능 서비스 이용자 보호 가이드라인 개발 과정	63

제1절 가이드라인 개발의 필요성 및 목적	63
1. 이용자 보호의 의미와 생성형 인공지능 시스템의 특수성	63
2. 제도적 대응 사례를 통해 본 가이드라인 개발의 필요	65
제2절 가이드라인 개발을 위한 연구반 구성 및 운영	72
1. 제1차 가이드라인 연구반 전체회의(kick-off)	73
2. 제2차 가이드라인 연구반 전체회의	75
3. 제3차 가이드라인 연구반 전체회의	83
4. 제4차 가이드라인 연구반 전체회의	90
5. 제5차 가이드라인 연구반 전체회의	93
6. 제6차 가이드라인 연구반 전체회의	94
제3절 가이드라인(안)에 대한 사업자 의견수렴	95
1. 사업자 현황 조사 및 자율규범 방향에 대한 의견수렴	96
2. 초안에 대한 사업자 간담회	99
3. 수정안에 대한 사업자 의견수렴	105
제4절 가이드라인 주요 주제 구성안	108
1. 주제1 이용자 인격권 침해	108
2. 주제2 결정 과정의 불투명성	110
3. 주제3 다양성 침해	111
4. 주제4 학습데이터 수집·활용 과정에서의 문제	113
5. 주제5 인공지능 생성 콘텐츠 활용에서 문제 발생 시 책임 소재의 불분명	114
6. 주제6 인공지능이 생성한 콘텐츠 유통에서 발생할 수 있는 문제	116
제5절 생성형 인공지능 서비스 이용자 보호 가이드라인 최종(안)	117
1. 생성형 인공지능 서비스 이용자 보호 가이드라인 최종(안)	118
2. 주제별 실천사례 예시(안)	122

제4장 알고리즘 추천서비스 규제 이행 실태 연구 **136**

제1절 연구 배경 및 중요성	136
1. 연구 배경	136

2. 기본원칙 현황	138
3. 연구의 중요성	141
제2절 연구의 목적	142
제3절 연구 내용 및 방법	143
1. 개요	143
2. 문헌 분석	144
3. 정책 및 서비스 현황 분석	146
4. 설문조사	147
5. 심층 인터뷰	149
제4절 연구 결과	149
1. 문헌 분석	149
2. 정책 및 서비스 현황 분석	154
3. 설문조사	173
4. 심층 인터뷰	212
제5절 기대효과	221
1. 미디어 플랫폼의 이용자 보호 원칙 준수 여부 평가	221
2. 설문조사 및 심층인터뷰 기반 정책적 시사점 도출	222
3. 이용자 권익 보호와 미디어 다양성 확보에 기여	222
4. 미디어 플랫폼의 신뢰성 제고에 기여	223
제5장 메타버스 및 알고리즘 사업자 규제 이행 실태 연구	224
제1절 연구 배경 및 목적	224
1. 연구 배경	224
2. 연구 목적	227
3. 기대효과	227
제2절 메타버스 서비스 활용 사례	228
1. 메타버스에 대한 정부의 제도적 지원	228
2. 메타버스 활용 사례	232

제3절 메타버스 이용자 보호 기본원칙 검토	239
1. 메타버스 이용자 보호 기본원칙 검토	239
2. 이용자 보호 기본원칙 검토	243
제4절 메타버스 사업자의 이용자 보호 실태 조사	244
1. 사업자 A	244
2. 사업자 B	251
제5절 시사점	261
1. 설문조사 결과에 대한 논의	261
2. 인터뷰 결과에 대한 논의	264
제6장 결 론	266
참고문헌	270

표 목 차

〈표 1-1〉	지능정보서비스 기본 원칙	2
〈표 2-1〉	방송통신위원회-NIA의 「생성형 AI윤리 가이드북」의 주요 쟁점 5가지	9
〈표 2-2〉	서울시-서울디지털재단의 「생성형 AI 윤리 가이드라인」 적용 대상 정의 ..	11
〈표 2-3〉	서울시-서울디지털재단의 「생성형 AI 윤리 가이드라인」 9대 핵심 요건 ...	12
〈표 2-4〉	국가정보원-국가보안기술연구소 「챗GPT 등 생성형 AI 활용 보안 가이드라인」 구성	13
〈표 2-5〉	「챗GPT 활용방법 및 주의사항 안내」의 챗GPT 공공 활용 분야	15
〈표 2-6〉	「챗GPT 활용방법 및 주의사항 안내」에서의 챗GPT 이슈와 활용 시 주의사항	15
〈표 2-7〉	국내 주요 대학의 생성형 인공지능 활용 가이드라인	16
〈표 2-8〉	해외 주요 대학의 인공지능 활용 가이드라인 사례	20
〈표 3-1〉	생성형 인공지능 서비스 이용자 보호 가이드라인 연구반 위원 명단	72
〈표 3-2〉	2024년 생성형 인공지능 서비스 이용자 보호 가이드라인 연구반 개최 현황	73
〈표 3-3〉	제2차 가이드라인 연구반 전체회의 프로그램	76
〈표 3-4〉	방통위의 이용자 보호를 위한 자율 규제	76
〈표 3-5〉	범정부 인공지능 윤리기준	78
〈표 3-6〉	국내 기업의 인공지능 윤리기준	79
〈표 3-7〉	기존 생성형 인공지능 가이드라인에서의 이용자 개념과 주요 내용	80
〈표 3-8〉	국토교통부 자율주행자동차 윤리 가이드라인	81
〈표 3-9〉	NIA 지능정보사회 윤리 가이드라인과 윤리현장	82
〈표 3-10〉	교육부 교육분야 인공지능 윤리 원칙	82
〈표 3-11〉	국가인권위원회 인공지능 개발과 활용에 관한 인권 가이드라인	83

〈표 3-12〉 주제1. 타인의 권리 및 사생활 침해(악의적 활용)	84
〈표 3-13〉 주제2. 투명성 및 설명가능성	85
〈표 3-14〉 주제3. 저작권 침해	86
〈표 3-15〉 주제4. 이용자 선택권 보장	87
〈표 3-16〉 주제5. 책임소재 불분명	87
〈표 3-17〉 주제6. 불법 콘텐츠 유통	88
〈표 3-18〉 주제7. 인간의 존엄성 훼손	88
〈표 3-19〉 주제8. 다양성 존중	88
〈표 3-20〉 주제9. 고지의무	89
〈표 3-21〉 가이드라인 초안에서 논의된 문제점	91
〈표 3-22〉 가이드라인 초안의 문제점을 해결하기 위한 개선방안 논의	92
〈표 3-23〉 사업자 의견수렴: 서비스 및 이용자 보호 조치에 대한 현황 조사	96
〈표 3-24〉 사업자 대면 간담회 일정	100
〈표 3-25〉 1차 사업자 간담회 주요 내용(네이버, 카카오, AWS, 메타 참여)	101
〈표 3-26〉 2차 사업자 간담회 주요 내용(뤼튼 테크놀로지스, 스캐터랩)	103
〈표 3-27〉 통신사 사업자 의견 수렴 주요 내용	104
〈표 3-28〉 글로벌 사업자를 포함하여 사업자 최종 의견수렴 주요 내용	106
〈표 3-29〉 인격권 보호의 실천사례: 미성년자 보호 조치	124
〈표 3-30〉 위험관리 전담 책임자 및 내부 조직 구성의 실천사례	129
〈표 3-31〉 생성형 인공지능 학습데이터 활용 여부의 안내 사례	131
〈표 3-32〉 이용약관에 이용자 오남용에 대한 책임 명시 사례	132
〈표 3-33〉 생성형 인공지능 서비스 이용 과정에서 이용자 책임 명시 경로	133
〈표 3-34〉 서비스 실행화면에 산출물의 부정확성을 안내하는 사례	134
〈표 3-35〉 이용자의 오남용에 대한 답변 거부완곡 응답 사례	135
〈표 5-1〉 국가별 주요 정책 및 지원 내용	232
〈표 5-2〉 분야별 메타버스 활용 사례	238
〈표 5-3〉 공동체 가치 형성 관련 조사 내용-사업자 A	244
〈표 5-4〉 이용자 참여 보장 관련 조사 내용-사업자 A	245

〈표 5-5〉 이용자 갈등 조정 관련 조사 내용-사업자 A	246
〈표 5-6〉 공정한 경제활동 보호 관련 조사 내용-사업자 A	248
〈표 5-7〉 데이터통제권 보장 관련 조사 내용-사업자 A	249
〈표 5-8〉 지속가능발전 지향 관련 조사 내용-사업자 A	250
〈표 5-9〉 정부역할 관련 조사 내용-사업자 A	251
〈표 5-10〉 공동체 가치 형성 관련 조사 내용-사업자 B	252
〈표 5-11〉 이용자 참여 보장 관련 조사 내용-사업자 B	252
〈표 5-12〉 이용자 갈등 조정 관련 조사 내용-사업자 B	253
〈표 5-13〉 공정한 경제활동 보호 관련 조사 내용-사업자 B	255
〈표 5-14〉 데이터통제권 보장 관련 조사 내용-사업자 B	256
〈표 5-15〉 지속가능발전 지향 관련 조사 내용-사업자 B	258
〈표 5-16〉 정부역할 관련 조사 내용-사업자 B	260

그 립 목 차

[그림 3-1] 인격권 보호의 실천사례: 경고문구 제시와 이용자 책임 범위 명시	123
[그림 3-2] 생성형 인공지능 산출물임을 알리는 워터마킹 적용 사례	125
[그림 3-3] 텍스트 기반 생성형 인공지능 산출물에서 정보 제공 사례	126
[그림 3-4] 기업 자체 인공지능 윤리 원칙 실행 사례	127
[그림 3-5] 서비스 화면에서 콘텐츠 신고 절차 사례	128
[그림 3-6] 사전·사후 생성형 인공지능 학습데이터 활용 여부 선택 사례	131
[그림 4-1] 인공지능 기반 미디어 추천서비스 이용자보호 핵심 원칙	139
[그림 4-2] 인공지능 기반 미디어 추천서비스 이용자보호 실행 원칙	141
[그림 4-3] 연구개발의 내용 및 방법	144
[그림 4-4] (A)기업 투명성 문항별 세부 점수	174
[그림 4-5] (B)기업 투명성 문항별 세부 점수	176
[그림 4-6] (A)기업과 (B)기업의 투명성 문항별 세부 점수 비교	177
[그림 4-7] (A)기업 공정성 문항별 세부 점수	179
[그림 4-8] (B)기업 공정성 문항별 세부 점수	180
[그림 4-9] (A)기업과 (B)기업의 공정성 문항별 세부 점수 비교	181
[그림 4-10] (A)기업 책무성 문항별 세부 점수	182
[그림 4-11] (B)기업 책무성 문항별 세부 점수	183
[그림 4-12] (A)기업과 (B)기업의 책무성 문항별 세부 점수 비교	184
[그림 4-13] (A)기업 핵심 원칙 항목별 평균 비교	186
[그림 4-14] (B)기업 핵심 원칙 항목별 평균 비교	187
[그림 4-15] 핵심 원칙 항목별 (A)기업과 (B)기업의 평균 비교	188
[그림 4-16] (A)기업 이용자 선택권 문항별 세부 점수	189
[그림 4-17] (B)기업 이용자 선택권 문항별 세부 점수	190
[그림 4-18] (A)기업과 (B)기업의 선택권 문항별 세부 점수 비교	191

[그림 4-19] (A)기업 자율검증 문항별 세부 점수	192
[그림 4-20] (B)기업 자율검증 문항별 세부 점수	193
[그림 4-21] (A)기업과 (B)기업의 자율검증 문항별 세부 점수 비교	194
[그림 4-22] (A)기업 불만 처리 및 분쟁 해결 문항별 세부 점수	195
[그림 4-23] (B)기업 불만 처리 및 분쟁 해결 문항별 세부 점수	196
[그림 4-24] (A)기업과 (B)기업의 불만처리 및 분쟁해결 문항별 세부 점수 비교	197
[그림 4-25] (A)기업 내부 규칙 문항별 세부 점수	198
[그림 4-26] (B)기업 내부 규칙 문항별 세부 점수	199
[그림 4-27] (A)기업과 (B)기업의 내부 규칙 문항별 세부 점수 비교	200
[그림 4-28] (A)기업 실행 원칙 문항별 평균 비교	202
[그림 4-29] (B)기업 실행 원칙 문항별 평균 비교	203
[그림 4-30] 실행 원칙 항목별 (A)기업과 (B)기업의 평균 비교	204
[그림 4-31] (A)기업 정부 역할 문항별 세부 점수	205
[그림 4-32] (B)기업 정부 역할 문항별 세부 점수	206
[그림 4-33] (A)기업과 (B)기업 정부 역할 문항별 세부 점수 비교	208
[그림 4-34] (A)기업과 (B)기업 서비스 제공자 지원(H1) 문항별 세부 점수 비교	209
[그림 4-35] (A)기업과 (B)기업 이용자 역량 강화 및 보호(H2) 문항별 세부 점수 비교	210
[그림 4-36] 정부 역할 항목별 (A)기업과 (B)기업의 평균 비교	211
[그림 5-1] 연구의 개요	226
[그림 5-2] 메타버스 이용자 보호 기본원칙	239

요 약 문

본 연구는 급속한 디지털 기술 발전으로 인한 방송통신 미디어 환경 변화에서 방송통신위원회가 추진해 온 이용자 보호 정책의 후속으로 생성형 인공지능 서비스 이용자 보호 가이드라인을 개발하는 과정을 기록하는 한편, 2019년 「이용자 중심의 지능정보사회를 위한 원칙」을 기조로 하여 개발된 2021년 「인공지능 기반 미디어 추천 서비스 이용자 보호 기본 원칙」과 2023년 「메타버스 이용자 기본 원칙」이 자율규범으로서 잘 활용되고 있는지를 점검한 결과를 담았다. 이는 인공지능이 일상의 디지털 기기 및 서비스에 내재화되는 과정에서 발생 가능한 이용자 권익 침해 이슈를 시의적절하게 점검, 대응하고, 법·제도가 정립되는 과정까지 이용자 보호를 위한 체계를 마련할 목적으로 수행되었다.

본 보고서는 2019년 방송통신위원회에서 발표한 「이용자 중심의 지능정보사회를 위한 원칙」 이후 후속적으로 추진된 정책개발 과정과 성과를 담았다.

먼저, 2장에서는 연구반 운영을 시작하면서 함께 조사, 분석한 자료를 정리하였다. 첫째, 기존에 개발된 생성형 인공지능 서비스 가이드라인들을 조사, 분석하였으며, 둘째 관련 규제 동향을 살펴보고 시사점을 도출하였다. 셋째 이용자의 니즈를 파악할 수 있는 실증연구 동향을 고찰하였으며 그 결과로부터 시사점을 도출하였다.

다음으로 제3장에서는 생성형 인공지능 서비스가 급속하게 일상 서비스로 확산하면서 초래되는 혼란이 이용자의 권익 침해로 이어지는 것에 예의 주시하고 모든 서비스 제공자가 공통으로 실천할 수 있는 실행 지침 개발을 제안하였다. 먼저 가이드라인 개발의 필요성과 목적, 이용자 범위와 정의, 생성형 인공지능의 특수성 등 연구반에서 논의한 여러 주제의 내용들을 간략하게 살펴보고, 다음으로 전체회의의 일정과 회의록도 간략히 소개하여 전체 논의 과정을 그려볼 수 있게 정리하였다. 또한 가이드라인 초안이 개발되기 전에 서비스 상황을 파악하기 위한 서면 조사 및 사업자 의견수렴 과정 등을 기록하였다. 나아가 문헌조사와 연구반 논의, 그리고 이해관계자 의견수렴 등을 거쳐 도출한 가이드라인(안)을 소개한다.

마지막으로 제4장 및 제5장에서는 인공지능 기반 미디어 추천 서비스 이용자 보호 기본 원칙」(19) 및 「메타버스 이용자 보호 기본원칙」(23) 수립 이후 기업 적용 현황 파악 및 인공지능 기술 발전과 이용 환경 변화에 따른 수정 및 보완 사항을 진단하였다. 이를 통해 자율 규범이 인공지능 서비스가 안정화되고 보편화되는 기간 동안에 어느 정도 효율적으로 작동할 수 있는지를 가늠해 볼 수 있을 것이다. 또한 이용자들이 인공지능 기반 추천서비스 및 메타버스 서비스를 안전하고 편리하게 이용할 수 있도록 산업 및 서비스 제공 환경을 장려하는 정책 후속 조치를 고려해 볼 근거가 될 수 있을 것으로 기대한다.

제1장 서론

제1절 연구의 필요성 및 목적

본 보고서는 2019년 방송통신위원회에서 발표한 「이용자 중심의 지능정보사회를 위한 원칙」 이후 후속적으로 추진된 정책개발 과정을 담은 보고서이다. 크게 두 주제의 정책개발 보고서로 구성되는데, 첫째 「생성형 인공지능 서비스 이용자 보호 가이드라인」을 개발하는 과정을 담고 있다. 둘째, 2019년 기본 원칙을 기반으로 인공지능 미디어 추천서비스의 특수성을 고려하여 2021년 개발된 「인공지능 기반 미디어 추천 서비스 이용자 보호 기본 원칙」(이하 '추천 서비스 기본 원칙')과 2023년 메타버스 플랫폼 서비스 제공자에게 권고되는 「메타버스 이용자 기본 원칙」이 잘 실천되고 있는지, 또 최근 급속한 기술 발전과 함께 보완될 요소가 있는지 등을 점검하는 보고서이다.

1. 방송통신위원회의 기본 원칙

2019년 방송통신위원회는 「이용자 중심의 지능정보사회를 위한 원칙」을 공표하였다. 이는 당시 인공지능을 비롯하여 빅데이터, 사물인터넷 등 여러 디지털 혁신기술이 방송통신 서비스와 결합하면서 이용자의 미디어 이용에 영향을 미치는 과정에서 기술적, 사회적 위험 요소를 최소화할 수 있는 장치를 마련하고자 하는 정책적 노력의 하나였다. 또한 이 원칙은 향후 빠르게 발전할 디지털 기술 및 미디어 환경 변화에 따라 서비스 제공자와 이용자 상호 간의 역할 재정립, 그리고 새로운 서비스 출현에서 제도적 미비나 산업적 우선순위로 인해 이용자의 기본권이 침해되는 상황을 방지할 목적으로 추진되었다. 해당 원칙에서 밝힌 원칙 작성의 목적은 다음과 같다.

“신기술의 도입이 초래할 수 있는 기술적, 사회적 위험으로부터 안전한 지능정보서비스 환경을 조성하기 위해, 지능정보사회의 모든 구성원들이 고려할 공동의 기본 원

칙을 제시한다.”

이 지능정보서비스 기본 원칙은 7개의 원칙으로 구성되는데, ① 사람 중심의 서비스 제공 원칙 ② 투명성과 설명가능성 원칙 ③ 책임성 원칙 ④ 안전성 원칙 ⑤ 차별 금지 원칙 ⑥ 참여 원칙 ⑦ 프라이버시와 데이터 거버넌스이다.

〈표 1-1〉 지능정보서비스 기본 원칙

기본 원칙	내용
사람 중심의 서비스 제공 원칙	- 지능정보서비스의 제공과 이용은 사람을 중심으로 그 기본적 자유와 권리를 보장하고 인간의 존엄성을 보호할 수 있는 방향으로 이루어져야 한다. - 지능정보사회의 구성원들은 각자의 자율성을 최대한 보장받으면서, 안전하고, 투명하며, 모든 사람이 고루 혜택을 누릴 수 있는 서비스가 이루어지는 환경을 조성하도록 노력한다.
투명성과 설명가능성 원칙	- 지능정보서비스에 포함된 서비스 체계와 작동방식이 이용자에게 중대한 영향을 끼칠 경우, 기업의 정당한 이익을 침해하지 않는 범위에서 이용자가 이해할 수 있도록 관련 정보를 작성한다. - 지능정보서비스가 인간의 신체, 자유, 재산 및 기타 기본권에 피해를 유발하였을 때, 기술적으로 가능한 범위에서 이용자에게 예측, 추천 및 결정의 기초로 사용한 시스템상의 주요 요인을 설명할 수 있어야 한다.
책임성 원칙	- 지능정보사회의 구성원들은 지능정보서비스의 올바른 기능과 사람 중심 가치의 보장을 위한 공동의 책임을 인식하고 관련한 법령과 계약을 준수한다. - 서비스의 제공과 이용으로 인해 발생한 타인의 권익 침해에 대한 법적, 사회적, 윤리적 책임을 다할 수 있도록 지능정보사회 구성원들은 지속적인 의견 교환에 참여한다.
안전성 원칙	- 안전하고 신뢰 가능한 지능정보서비스의 개발과 이용을 위해 지능정보 사회 참여자 모두가 노력한다. - 지능정보사회가 초래할 수 있는 피해에 대한 복구 방안을 사전에 확보하기 위한 노력으로서, 제공자와 이용자는 상호 협의 하에 그에 대한 자율적인 대비 체계를 수립하고 운영한다.
차별 금지 원칙	지능정보사회의 구성원들은 지능정보서비스가 사회적, 경제적 불공평이나 격차를 초래할 수 있다는 점을 인식하고, 기술 사용에 있어 사회적 다양성을 고려한다.

기본 원칙	내용
차별 금지 원칙	알고리즘 설계, 데이터의 수집과 입력 및 알고리즘 실행 등 개발과 사용의 모든 단계에서 차별적 요소를 최소화할 수 있도록 노력한다.
참여 원칙	- 지능정보사회의 구성원들은 공적인 이용자 정책 과정에 차별 없이 참여할 수 있다. - 공적 주체는 지능정보서비스의 이용과 관련하여 제공자와 이용자가 실질적으로 의견을 제시할 수 있는 정기적인 통로를 조성하여야 한다.
프라이버시와 데이터 거버넌스	- 지능정보사회 구성원들은 서비스의 개발, 공급 및 이용의 전 과정에서 개인정보 및 프라이버시를 보호하며, 특히 지능정보서비스의 제공과 이용 과정에서 데이터 사용이 프라이버시에 미치는 부정적인 영향을 최소화하도록 노력한다. - 지능정보사회 구성원들은 데이터의 활용을 통한 기술적 이익의 향유와 프라이버시 보호 사이의 균형을 위해, 지속적인 의견 교환에 참여한다.

방송통신위원회는 이 기본 원칙을 기반으로 하면서 기술 발전과 새로운 방송통신 미디어 환경 변화를 반영하여 2021년 「인공지능 기반 미디어 추천 서비스 이용자 보호 기본 원칙」, 2023년 「메타버스 이용자 기본 원칙」을 발표하였다. 2021년 ‘추천 서비스 기본 원칙’은 디지털 미디어가 플랫폼 중심으로 내용과 형식이 재편되는 추세에 따라 미디어 이용자 권익 보호를 위한 새로운 규율이 필요한 시점에 개발되었다. 이 기본 원칙은 특히 디지털 미디어 플랫폼에서 제공되는 인공지능 기반 추천 서비스의 불투명성 및 편향성 등의 이슈와 관련하여 이용자의 정보접근권과 선택권을 보장하는 자율 규범 형식이다. 주요 내용은 추천서비스 제공자가 실현해야 할 3대 핵심 원칙과 5대 실행 원칙으로 구성되었다. 3대 핵심 원칙으로는 ① 투명성, ② 공정성, ③ 책무성이 제시되었으며, 5대 실행 원칙에는 ① 이용자를 위한 정보 공개, ② 이용자의 선택권 보장, ③ 자율검증 실행, ④ 불만 처리 및 분쟁 해결, ⑤ 내부 규칙 제정이 포함된다. 이 가운데 ‘정보 공개’와 ‘선택권 보장’은 이용자에게 추천 서비스에 관한 기본 정보와 개인 설정 기능을 제공함으로써 이용자의 합리적 소비를 유도하는 동시에 추천 서비스가 실제 선호와 필요에 따라 실행되도록 하여, 추천 서비스의 본질적 기능을 보완하는 의미를 갖는다고 밝히고 있다.

2023년 발표된 「메타버스 이용자 기본 원칙」은 최근 몇 년 전부터 가상과 현실을 연계하는 세상인 메타버스(초월의 Meta + 세상의 Universe)가 새롭게 등장하여 아바타를 매개

체로 하여 사업자들에게는 새로운 경제적 이익 창출 기회를 제공하고, 이용자들에게는 실재감과 몰입감 높은 경험을 제공하는 서비스 환경에서 발생할 수 있는 이용자 권익 침해 요소를 관리하는 자율규제로 발표되었다. 이 기본 원칙은 “메타버스 플랫폼 서비스에서 민주적 공동체, 신뢰 기반 환경, 경제적 기회 보장의 원칙을 확립하여 모든 디지털 시민이 안전하게 편익과 권리를 누리는 공동체 발전을 도모하기 위해 ‘메타버스 플랫폼 서비스 제공자’에게 권고되는 자율적인 실천규범”이며, 6대 기본 원칙과 5개 실천 규범으로 구성되었다. 6대 기본 원칙은 ① 공동체 가치 형성, ② 이용자 참여 보장, ③ 이용자 갈등 조정, ④ 공정한 경제활동 보호, ⑤ 데이터 통제권 보장, ⑥ 지속가능발전 지향이다.

2. 가이드라인 개발의 필요: 생성형 인공지능 서비스의 확산과 이용자 권익 침해 이슈

방송통신위원회는 생성형 인공지능 서비스가 급속하게 일상 서비스로 확산하면서 초래되는 혼란이 이용자의 권익 침해로 이어지는 것에 예의 주시하고 모든 서비스 제공자가 공통으로 실천할 수 있는 실행 지침 개발을 제안하였다.

앞서 소개한 「이용자 중심의 지능정보사회를 위한 원칙」(‘19)이나 ‘추천 서비스 기본 원칙’(21)에서도 확인할 수 있듯이, 인공지능 기술의 특수성으로 인해 발생할 수밖에 없는 이슈들이 있다. 먼저 인공지능 기술의 특성상 데이터 학습 과정을 필수로 거치는데 이 과정에서 과거 사회가 축적한 편견이 데이터를 통해 인공지능의 산출물에 반영되는 문제가 있다. 이는 다양성 존중을 해하거나 프라이버시 침해와 같은 기본권과도 연결될 수 있는 이슈이다. 또 인공지능의 추론 과정 등 인간이 알 수 없고 통제할 수 없는 영역이 존재한다. 이는 결과물의 공정성을 신뢰할 수 없거나 불명확한 책임소재 등과 관련된다. 또 인공지능 모델 개발 단계에서 특정 이해관계만을 위해 알고리즘이 개발된다면 인간 존엄성을 해하고 차별을 정례화할 위험이 발생할 수 있다. 이에 투명성과 설명가능성, 인간 존엄과 차별 금지의 원칙이 강조된다. 또 새로운 기술을 수용하는 과정에서 발생할 수 있는 위험 요소를 기술적, 제도적으로 관리해야 한다는 차원에서 안전의 이슈도 기본적으로 제기된다.

이에 더해 생성형 인공지능의 특수성으로 인한 추가 이슈를 생각해 볼 수 있다. 예를 들어, 정보의 부정확성 이슈이다. 기존 인공지능 모델이 특정 문제를 해결하기 위해 구조화된 데이터를 지도 학습 형태로 학습하였다면, 생성형 인공지능 콘텐츠 생성을 목적으로 대규모의 비구조화된 데이터를 사전 학습하여 인간의 창의적 작업을 모방한 결과물을 내

놓는다. 콘텐츠 생성에 목적을 두다 보니 정확성을 보장하지 않는다. 이러한 기술적 특성에도 불구하고 챗GPT가 처음 선보일 때 새로운 유형의 정보 검색 서비스로 소개되었기에 이용자들은 인공지능 산출물의 부정확한 정보에 대한 이해와 대응력이 부족한 상태에서 많은 혼란이 있었다.

둘째, 생성형 인공지능이 산출하는 결과물이 인간의 창작물과 유사하다는 데서도 문제가 발생할 수 있다. 생성형 인공지능은 대규모 데이터를 학습해 새로운 콘텐츠를 생성하는 것이 주요 기능이며, 특히 언어 영역에서는 인간 수준으로 단어 순서와 문맥을 효과적으로 이해하고 문장 생성이나 번역, 요약 등에 뛰어나다. 이러한 능력은 정보의 정확성보다는 ‘그럴듯한 콘텐츠 생성’에 더 목적을 둔다.

지금 시점에서는 해당 디지털 콘텐츠가 인간이 생성한 것인지 인공지능이 생성한 것인지 인지하는 것이 중요하다. 예를 들어, 인간의 창작물은 법적 보호를 받지만, 인공지능 생성물은 그러한 권리가 없기 때문에 인간의 창작물인지 인공지능의 생성물인지 분명히 할 필요가 있다. 또 허위 정보 여부를 판단하거나 인공지능과의 상호작용에 지나치게 몰입하는 것을 방지하는 데에도 인공지능 생성물이라는 것을 인지할 필요가 있다. 무엇보다 해당 콘텐츠가 누구에 의해 어떤 과정과 목적으로 생성되었는지를 알 때, 이용자는 그 콘텐츠에 대해 좀 더 정확한 판단을 할 수 있다. 이처럼 생성형 인공지능이 만들어 낸 콘텐츠임을 사전에 인지할 수 있게 함으로 해서 여러 문제의 발생 가능성을 감소시키는 효과를 얻는다.

셋째, 생성형 인공지능의 산출물은 이용자의 상호작용 과정에서 영향을 받는다. 예를 들어, 특정 인물에 대한 정보를 얻고자 질문을 프롬프트에 입력할 때, 이용자가 어떻게 입력하느냐에 따라 그 결과에 차이가 있다. 이러한 특성으로 인해 생성형 인공지능 서비스를 사용하면서 발생할 수 있는 이슈에 대해 이용자에게 책임을 묻는 경우가 많다. 이용자가 어떤 입력 프롬프트값을 넣느냐에 따라 해당 위험이 발생할 수도 발생하지 않을 수도 있기 때문이다. 그러나 다른 한편에서는 신기술 및 신서비스 론칭 초기 단계에서, 즉 생성형 인공지능의 기술적 특성에 대한 이해도 부족하고 제도적 보호 요소도 부재한 상태에서 이용자에게 과도한 책임을 지우는 것은 아닌지에 대한 문제를 제기할 수 있다.

이 인공지능 및 생성형 인공지능 서비스의 기술적 특수성은 이 가이드라인을 개발하는 이유이자 내용 구성의 주요 기준이 되었다. 즉, 본 가이드라인을 개발할 때, 생성형 인공지능

능의 기술적 구조적 특성 및 이용 행태를 고려하여 효과적인 이용자 보호를 위한 자율규제를 마련하고자 하였으며, 본 보고서에 그 과정을 충실하게 담았다.

3. 자율규범의 충실한 이행: 사업자의 기본원칙 준수 상황 점검

본 보고서의 두 번째 주요 목표는 인공지능 기반 미디어 추천 서비스 이용자 보호 기본 원칙¹⁹⁾ 및 「메타버스 이용자 보호 기본원칙」²³⁾ 수립 이후 기업 적용 현황 파악 및 인공지능 기술 발전과 이용 환경 변화에 따른 수정 및 보완 사항을 진단해 보는 것이다. 이를 통해 자율 규범이 인공지능 서비스가 안정화되고 보편화되는 기간 동안에 어느 정도 효율적으로 작동할 수 있는지를 가늠해 볼 수 있을 것이다. 또한 이용자들이 인공지능 기반 추천서비스 및 메타버스 서비스를 안전하고 편리하게 이용할 수 있도록 산업 및 서비스 제공 환경을 장려하는 정책 후속 조치를 고려해 볼 근거가 될 수 있을 것이다.

인공지능 기술은 혁신성과 편리함으로 인해 방송·통신 환경을 획기적으로 변화시킬 기술로 주목받는 한편, 여전히 끊임없이 빠르게 발전하고 있고 그만큼 불확실성과 불완전성이 있는 기술로서 어느 정도 통제가 필요하다는 평가를 받는다. 오랫동안 인간의 고유 영역이었던 ‘사고’와 ‘판단’ 영역을 지원하는 전례 없는 혁신 도구이다. 전례가 없기 때문에 잠재적 위험에 대한 사회제도적 대응이 부재하며, 실증 사례가 충분하지 않기에 법적 대응도 어렵다. 이 시점에서는 생태계 내 주요 주체들이 잠재적 위험에 대해 자율적으로 규범을 만들어 실천하는 사회윤리에 의존할 수밖에 없다. 이러한 사회의 필요에 따라 건전한 방송·통신 미디어 환경조성을 주 목표로 하는 방송통신위원회가 이해관계자와의 협의를 통해 「인공지능 기반 미디어 추천 서비스 이용자 보호 기본 원칙」²¹⁾을 마련하였으며, 3년이 지난 시점에서 그 효용성을 재점검하는 것은 타당하다.

인공지능 못지않게 미래사회 방송·통신 환경에 영향을 미치는 디지털 기술 환경으로 ‘메타버스’에 주목해 왔고, 그에 따라 2023년 「메타버스 이용자 기본 원칙」을 개발하여 역시 새로운 환경 변화에 우선 자율 규범으로 대응하고자 하였다. 아직 산업 및 서비스 생태계에 공유된 지 1년이 채 되지 않은 시점에서 자율규범 이행 여부와 효용성을 점검하기에는 시기상조이지만, 그 방법론이 인공지능 기반 미디어 추천 서비스의 기본원칙과 크게 다르지 않을 것이라는 판단 하에 병행하였다. 이번의 경험과 성과는 기본 원칙 발표 이후

3년이 되는 시점에서 더 체계적인 자율규범 효용성 점검과 평가의 실증 사례가 될 것으로 기대한다.

제2절 연구 주요 구성

본 보고서는 급속한 디지털 기술 발전으로 인한 방송통신 미디어 환경 변화에서 방송통신위원회가 추진해 온 이용자 보호 정책의 후속으로 생성형 인공지능 서비스 이용자 보호 가이드라인을 개발하는 과정을 기록하는 한편, 2019년 「이용자 중심의 지능정보사회를 위한 원칙」을 기조로 하여 개발된 2021년 「인공지능 기반 미디어 추천 서비스 이용자 보호 기본 원칙」과 2023년 「메타버스 이용자 기본 원칙」이 자율 규범으로서 잘 활용되고 있는지를 점검한 결과를 담았다. 이는 인공지능이 일상의 디지털 기기 및 서비스에 내재화되는 과정에서 발생 가능한 이용자 권익 침해 이슈를 시의 적절하게 점검, 대응하고, 법·제도가 정립되는 과정까지 이용자 보호를 위한 체계를 마련할 목적으로 수행되었다. 다음은 각 장의 주요 내용이다.

먼저 2장과 3장은 「생성형 인공지능 서비스 이용자 보호 가이드라인」을 개발하는 과정과 개발(안)을 소개한다.

2장에서는 「생성형 인공지능 서비스 이용자 보호 가이드라인」 개발에 앞서 살펴본 동향 조사뿐만 아니라 개발 과정에서 추가로 나온 국내외 법제 동향도 포함하여 정리하였다.

3장은 먼저 가이드라인 개발의 필요성과 목적, 이용자 범위와 정의, 생성형 인공지능의 특수성 등 연구반에서 논의한 여러 주제의 내용들을 간략하게 정리하였다. 다음으로 전체 회의 일정과 회의록도 간략히 소개하여 전체 논의 과정을 그려볼 수 있게 정리하였다. 전체 회의는 9명의 외부전문위원과 정보통신정책연구원의 연구진 4명, 그리고 방송통신위원회 3명 등 총 16명이 참여하는 회의였다. 참고로 전체회의에서 논의된 내용을 기반으로 텍스트화 작업을 하는 실무회의가 따로 7회 진행되었다. 마지막으로 사업자 의견수렴 과정을 기록하였다. 사업자 의견수렴은 가이드라인 초안이 개발되기 전에 서비스 상황을 파악하기 위한 서면 조사를 2월~4월까지 수행하였다. 수행 기간이 길어진 이유는 글로벌 기업의 경우, 한글 질문지를 영문으로 번역하여 미국 본사로 보내고 미국 본사에서 검토 및 답변 후 다시 한국 지사로 보내져 번역 절차를 다시 밟는 등 다소 복잡한 단계를 거쳐야했기

때문이다. 또한 문헌조사와 연구반 논의, 그리고 이해관계자 의견수렴 등을 거쳐 도출한 가이드라인(안)을 소개한다.

다음 4장과 5장은 「인공지능 기반 미디어 추천 서비스 이용자 보호 기본 원칙」(21)과 「메타버스 이용자 기본 원칙」(23)이 자율규범에 얼마나 기여하는지를 점검하는 과정과 결과를 담았다. 구체적으로 4장에서는 인공지능 기반 미디어 추천 서비스의 이용자 보호를 위해 마련된 핵심 원칙과 실행 원칙의 이행 실태를 검증하기 위해 문헌 분석, 정책 및 사례 분석, 설문조사, 심층 인터뷰의 4가지 연구 방법을 종합적으로 활용하였다. 문헌 분석을 통해 이론적 토대를 마련하고, 정책 및 사례 분석을 통해 주요 플랫폼의 운영 정책과 서비스 현황을 비교·평가하였다. 설문조사를 통해 주요 미디어 플랫폼 사업자의 이행 수준을 정량적으로 측정하였으며, 심층 인터뷰를 통해 설문조사의 맥락을 보완하고 구체적인 운영 사례와 개선 방안을 청취하였다. 이를 바탕으로 각 원칙의 이행 수준을 체계적으로 평가하고, 주요 플랫폼의 운영 현황을 파악하며, 정책 개선을 위한 실질적 시사점을 도출하고자 하였다.

5장에서는 메타버스의 규제 이행 실태를 종합적으로 점검하여 이용자 보호 기본원칙의 실효성을 평가하고 정책적 시사점을 도출하고자 하였다. 이를 위해, 본 연구는 다음과 같은 세부 내용을 수행하였다. 먼저, 메타버스 플랫폼 서비스 제공자들의 이용자 보호 기본원칙 준수 여부를 체계적으로 조사하고 분석하였다. 이를 통해 현재의 자율규제 체계의 실효성을 평가하고, 개선이 필요한 영역을 식별하였다. 둘째, 이용자 보호 기본원칙의 실제 적용 과정에서 발생하는 문제점과 한계를 파악하여, 실효적 규제 방안을 모색하고자 하였다. 마지막으로, 연구 결과를 바탕으로 이용자 보호를 강화하고 디지털 환경의 신뢰성을 제고하기 위한 구체적인 정책 제언을 제시하였다.

마지막으로 6장에서는 두 정책연구의 성과를 요약, 정리하고 향후 과제를 제시한다.

제2장 생성형 인공지능 서비스 규제 동향과 이용자 연구

이번 장에서는 「생성형 인공지능 서비스 이용자 보호 가이드라인」 개발에 앞서 살펴본 동향 조사 내용과 이후 개발 과정에서 추가된 국내외 법제 동향도 포함하여 정리하였다.

제1절 생성형 인공지능 서비스 가이드라인 동향

1. 공공기관의 생성형 인공지능 활용 가이드라인

가. 방송통신위원회-NIA의 「생성형 AI윤리 가이드북」

먼저 방송통신위원회와 한국지능정보사회진흥원(NIA)에서 발간한 ‘생성형 AI윤리 가이드북’을 살펴본다. 2023년 12월 발간된 이 가이드북은 이용자들이 인공지능 기술로 인한 역기능에 대응할 수 있는 역량 증진을 목적으로 한다. 해당 가이드라인에는 이용자 스스로 판단과 검토를 통해 생성형 인공지능을 책임감 있게 이용할 수 있는 능력 및 윤리적 마인드를 갖추는 것으로 기술되어 있다. 이용자가 지침을 따르는 주요 행위자가 되며 실제로 생성형 인공지능을 이용하는 개인 이용자들이 주의해야 할 점들을 설명하고 있다.

〈표 2-1〉 방송통신위원회-NIA의 「생성형 AI윤리 가이드북」의 주요 쟁점 5가지

주요 쟁점	내용
저작권	AI 생성물의 저작권은 누구에게?
책임성	AI가 다 했으니 사람은 책임 없다?
허위조작정보	AI 챗봇으로 보이스 피싱까지?
개인정보·인격권	생성형 AI는 당신이 지난 여름에 한 일을 알고 있다?
오남용	생성형 AI는 나만의 일타 강사? 생성형 AI는 도구일뿐!

자료: 방송통신위원회-NIA(2023. 12). 「생성형 AI윤리 가이드북」

저작권, 책임성, 허위조작정보, 개인정보·인격권, 오남용 등 다섯 가지의 주요 주제를 다루며, 세부 이슈들은 질문과 답변의 형식으로 제공된다. 즉, 이용자들이 생성형 인공지능을 사용하며 마주치게 되는 문제에 대해 답변하는 방식으로 기술되어 있다. 각 주제의 주요 내용은 다음과 같다.

첫째, 저작권과 관련하여 일반적으로 생성형 인공지능의 저작권을 인정하고 있지 않음을 밝힌다. 두 번째 쟁점인 책임성과 관련하여 생성형 인공지능을 통해 만들어진 결과물을 이용하는 이용자의 책임을 강조한다. 즉, 이용자는 대학 과제나 포스터 공모전, 회사 보고서 작성 등 다양한 상황에서 생성형 인공지능을 활용한 사실을 밝히고 활용 시 발생하는 상황에 대한 책임 주체가 되어야 한다고 안내한다. 셋째, 허위조작정보에서는 생성형 인공지능을 이용하여 가짜뉴스를 생성하거나 보이스 피싱에 이용하는 등 다양한 악용 사례에 대한 이용자의 비판적 시각과 주의를 강조하고 있다. 넷째, 개인정보·인격권에 대해서는 생성형 인공지능 서비스에 개인정보, 기밀 등을 입력하지 않을 것을 당부한다. 마지막 오남용 사례에서는 생성형 인공지능에 과의존하거나 중독되지 않도록 경계하며, 인간 고유의 능력을 보존할 수 있도록 생성형 인공지능을 하나의 도구로써 사용하도록 권장하고 있다.

나. 서울시-서울디지털재단의 「생성형 AI 윤리 가이드라인」

서울시와 서울디지털재단은 지난 2023년 11월 서울시 공무원과 서울시민 등 서울시의 이해관계자 두루 준수할 수 있는 「생성형 AI 윤리 가이드라인」을 발간하였다. 주요 내용을 살펴보면, 먼저 생성형 인공지능과 전통적 인공지능 간의 차이점을 제시하고 이용자가 어떤 목적으로 어떻게 생성형 인공지능을 활용해야 하는지 안내하고 있다. 가이드라인의 주요 적용 범위는 서울시에 도입되는 생성형 인공지능의 개발과 활용으로 한정하였고, 적용대상을 “생성형 AI를 활용해 서울시의 서비스를 개발·활용하는 모든 이해관계자”로 정의한다. 흥미로운 점은 이해관계자인 이용자, 개발자, 운영자 간의 요건 중복이 있을 수 있음을 밝히고 있다.

〈표 2-2〉 서울시-서울디지털재단의 「생성형 AI 윤리 가이드라인」 적용 대상 정의

대상	정의
이용자	서울시의 생성형 인공지능을 활용하는 시민과 공무원이 주요 대상
개발자	서울시 생성형 인공지능을 기획·설계·구축하는 개발자를 총칭
운영자	서울시 생성형 인공지능을 도입하여 실질적으로 운영하는 공무원·유관기관 임직원을 의미

자료: 서울시-서울디지털재단(2023. 11). 「생성형AI 윤리 가이드라인」

가이드라인의 주요 내용은 9대 핵심요건을 중심으로 구성되어 있으며, 적용 대상 중 이용자에 대한 비중이 높다(〈표 2-3〉). 예를 들어, 생성형 인공지능의 허위정보 답변 이슈에서는 이용자의 점검/검토 의무를, 개인정보·중요정보 유출에서는 이용자가 프롬프트에 입력하지 않아야 할 사항을 안내한다. 또 이용자가 생성형 인공지능 보안이 충분하지 않을 수 있다는 사실에 주의할 것, 그리고 탄소 배출량 증가 가능성도 인지하도록 안내한다. 이처럼 해당 가이드라인은 생성형 인공지능 이용에 대한 이용자들의 주의와 대처에 비중을 두고 개발되었다.

〈표 2-3〉 서울시-서울디지털재단의 「생성형 AI 윤리 가이드라인」 9대 핵심 요건

이용자의 검증 역량 (허위정보 답변)	4. 사용가능성 확보 (결과물의 저작권 침해)	7. 공공성 (사회적 문제 야기)
▶ 이용자는 활용 상황에 따라 생성형 인공지능의 도출 정보를 검증하고 활용하는 역량을 확보 ▶ 이용자는 생성형 인공지능의 결과물을 객관성이 요구되는 상황에 활용할 경우 편향·차별 등의 내용이 포함되어 있는지 점검	▶ 이용자는 생성형 인공지능 결과물을 활용할 때 사용가능성을 확보하기 위해 노력 ▶ 이용자는 생성형 인공지능 결과물을 활용할 때 타인의 권리침해 여부를 법·제도적으로 고려	▶ 이용자는 생성형 인공지능을 보조적 수단으로 인지·활용해야 하며 공공성을 침해하지 않도록 노력 ▶ 운영자는 생성형 인공지능 도입 시 해당 서비스 관련 일자리에 미치는 영향을 고려
2. 책임성 (악의적 활용)	5. AI 학습데이터의 안전성 (저작권 침해와 오용·편향 데이터 학습)	8. 생태적 지속가능성 (탄소 多배출)
▶ 이용자는 생성형 인공지능을 개발 목적에 부합하도록 활용 ▶ 이용자는 생성형 인공지능 결과물이 사회 전반에 부정적인 영향을 미치지 않도록 노력 ▶ 이용자는 자신이 생성을 유도한 생성형 인공지능의 결과물에 대한 책임을 인지	▶ 개발자는 타인의 저작권을 침해하지 않은 범위에서 데이터를 학습 ▶ 개발자는 오용·편향 데이터를 학습에 활용하지 않아야 함	▶ 이용자·운영자는 생태적으로 설계된 서비스를 도입·활용하기 위해 노력 ▶ 개발자는 데이터 구축 단계에서 최소한의 탄소배출을 위해 노력
3. 데이터 권리와 책임 (개인정보·중요정보 유출)	6. AI 윤리 소양 (이용자 이해도 증진)	9. 보안성 (생성형 AI 자체의 보안)
▶ 이용자는 개인·중요정보가 생성형 인공지능에 노출되지 않도록 유의 ▶ 이용자는 생성형 인공지능을 활용함과 동시에 서비스가 이용자 정보를 수집한다고 인식 ▶ 이용자는 개인정보에 대한 책임과 권한이 자신에게 있음을 인식	▶ 이용자는 생성형 인공지능 특성을 이해하고 사용상의 문제점을 인식하려고 노력해야 하며, 운영자는 이용자의 인공지능을 윤리 소양을 가질 수 있도록 노력	▶ 이용자·운영자는 생성형 인공지능이 가진 해킹 등 보안 위험에 유념 ▶ 개발자는 생성형 인공지능에 대한 보안 사항을 적시에 설명하고 통보 ▶ 이용자는 생성형 인공지능으로 인해 보안 상 피해를 입은 경우 책임과 보호조치를 요구해야 하며, 운영자·개발자는 보안사고 발생 시 명확한 책임 주체를 설정하고 통보

자료: 서울시 생성형AI 윤리 가이드라인(2023. 11)

다. 국가정보원-국가보안기술연구소 「챗GPT 등 생성형 AI 활용 보안 가이드라인」

국가정보원과 국가보안기술연구소는 2023년 6월 29일에 챗GPT 관련 보안 문제를 예방하기 위해 「챗GPT 등 생성형 AI 활용 보안 가이드라인」을 공개하였다. 해당 가이드라인은 환각현상, 데이터 유출, 플러그인·확장 프로그램·API 취약점 등 대표적인 생성형 인공지능 기술 보안 위협의 원인과 사례에 대해 설명함으로써, 각급기관의 '정보화·정보보안 담당자'들이 보다 효율적이고 안전하게 서비스를 이용할 수 있도록 하는데 목적을 두고 있다. 이 외에도 생성형 인공지능 서비스의 잠재적인 위험을 최소화하고, 올바른 활용을 통한 업무 효율성 향상을 도모하기 위함임을 밝히며, 안전하고 올바른 사용을 위한 방법을 문답 형식으로 제시한다. 총 5개의 주제로, ① 서비스 사용 주의사항, ② 서비스와의 대화 시 주의사항, ③ 인공지능 모델 플러그인 사용 주의사항, ④ 인공지능 모델 확장 프로그램 사용 주의사항, ⑤ 인공지능 모델 생성 기반 공격 대처 방안에 대해 안내하고 있으며, 자세한 내용은 아래 <표 2-4>와 같다.

<표 2-4> 국가정보원-국가보안기술연구소 「챗GPT 등 생성형 AI 활용 보안 가이드라인」 구성

주제	세부 내용
서비스 사용 주의사항	• 서비스 접근 • 계정 관리법
서비스와의 대화 시 주의사항	• 답변 검증(정확성 및 유해성) • 개인정보 및 민감 정보 처리 • 인공지능 모델이 생성한 데이터 관리 • 책임감 있는 사용 • 업무에서 올바르게 인공지능 모델 활용하기(이용목적 및 협업)
AI 모델 플러그인 사용 주의사항	• 올바른 서비스 플러그인 사용 및 관리 • 개인정보 및 민감 정보 처리
AI 모델 확장 프로그램 사용 주의사항	• 올바른 서비스 확장 프로그램 사용 및 관리 • 개인정보 및 민감 정보 처리
AI 모델 생성기반 공격 대처 방안	• 인공지능 모델 생성기반 공격 정의 • 인공지능 모델 생성기반 공격 대처

자료: 국가정보원-국가보안기술연구소(2023. 6). 「챗GPT 등 생성형 AI 활용 보안 가이드라인」

각 주제의 대표 사례를 소개하면, 서비스 사용 주의사항 중 서비스 접근의 경우, 공식 사이트를 이용하고, 안전한 네트워크 환경에서 이용하거나, 웹 브라우저 보안을 강화하는 등의 조치를 취할 것을 권장한다. 또 다른 예로, 서비스와의 대화 시 주의사항으로 책임감 있는 서비스 사용을 권장하고 있으며, 세부적으로는 윤리적 사용, 악의적 행위 금지, 서비스 약관 준수를 권하고 있다. 이 외에도 생성형 인공지능 기반 서비스 구축·활용을 위한 가이드라인 또한 제공하고 있으며, ① 구축 유형에 따른 인공지능 모델 도입 시 고려사항, ② 인공지능 모델 API 활용 및 민간 인공지능 모델 도입 방안, ③ 자체 데이터 세트 및 인공지능 모델 구축 단계별 보안 위협 대응 방안 등 실무자가 생성형 인공지능을 활용한 서비스 구축할 때 안전한 활용을 위한 절차를 설명한다.

라. 행정안전부 「챗GPT 활용방법 및 주의사항 안내」

행정안전부는 챗GPT 활용방법과 주의사항을 기재한 가이드라인을 2023년 5월에 배포하였다. 행정안전부의 보도자료에 따르면, 해당 가이드라인은 공무원들이 실제 업무와 관련하여 안전하게 이용할 수 있도록 개발되었으며, 공공부문에서 챗GPT(ChatGPT)를 올바르게 활용할 수 있도록 중앙행정기관과 지방자치단체 약 300개 기관에 이 안내서를 배포하였다.

본문에서는 공공부문의 주요 활용 영역을 ‘정보탐색능력 활용’, ‘언어능력 활용’, ‘컴퓨터 능력 활용’으로 나누고 세부 활용방법과 예시를 제공한다(표 2-5). 예를 들어, 보도자료 등 초안 작성을 위해 챗GPT를 사용하려고 할 때, ‘인공지능의 공공부문 도입을 주제로 하는 세미나에서 인사말을 해야하는데 초안을 1,500자 내외로 작성해주세요’와 같이 입력할 수 있음을 안내한다. 또, 챗GPT의 문제점과 활용 시 주의사항 또한 제공하여 이용자들의 안전한 사용을 장려하고 있다(표 2-6).

〈표 2-5〉 「챗GPT 활용방법 및 주의사항 안내」의 챗GPT 공공 활용 분야

분야	예시
정보탐색능력 활용	• 기획 보고서 작성을 위한 아이디어 탐색 • 업무에 필요한 국내외 자료 조사
언어능력 활용	• 보도자료, 인사말, 강의자료 등 대외 공개자료 초안 작성 • 언론기사, 논문, 보고서 등 외부자료 요약 • 해외사례 조사, 해외 홍보 등을 위한 언어 간 번역
컴퓨터능력 활용	• 엑셀 등 응용프로그램 사용법 • 업무 자동화 프로그램 코드 생성

자료: 행정안전부 챗GPT 활용방법 및 주의사항 안내(2023. 5)

〈표 2-6〉 「챗GPT 활용방법 및 주의사항 안내」에서의 챗GPT 이슈와 활용 시 주의사항

문제점	• 저작권·개인정보 보호 • 중요정보 유출 (예) OO직원이 프로그램 코드를 입력한 사실이 확인되었고 이용 제한 조치가 된 사례가 있음 • 답변의 신뢰성, 윤리성, 편향성
활용 시 주의사항	• 챗GPT에 비공개 정보나, 개인정보를 입력하지 말 것 (예) 외부 반출이 허용되지 않는 정보 - 국산 전투기 무기 개량 방향에 대한 보고서에 작성을 위해 현재 무기 사양 등을 입력 • 챗GPT가 생성한 답변을 사실여부 검증 없이 이용하지 말 것

자료: 행정안전부(2023. 5). 「챗GPT 활용방법 및 주의사항 안내」. 연구진 작성

2. 대학의 생성형 인공지능 활용 가이드라인

가. 국내 주요 대학의 생성형 인공지능 활용 가이드라인

생성형 인공지능 서비스의 선발주자인 챗GPT를 가장 빠르게 도입한 곳이 대학이다. 이에 국내외 주요 대학들은 학생뿐만 아니라 교직원, 교수 등 학교 내 모든 구성원이 챗GPT를 잘 활용할 수 있는 지침을 개발하여 배포하였다.

먼저 〈표 2-7〉 국내 주요 대학이 개발, 배포한 가이드라인의 주요 내용을 정리한 것이다. 대부분 생성형 인공지능을 안전하고 윤리적으로 활용할 때의 긍정 효과를 염두에 두고 각 이용 주체인 학습자, 교수자, 연구원, 교직원 등의 안전한 활용 기준을 제시한다. 예

를 들어, 활용 시 생성형 인공지능 답변의 진실성, 표절 가능성, 학문적 진실성이나 연구의 투명성을 저해할 위험성 등을 언급하면서 이에 대한 주의 사항을 안내하는 방식이다. 대학의 생성형 인공지능 활용 가이드라인은 더 나아가 대학 구성원들의 비판적 사고 능력 함양과 투명하고 윤리적인 사용을 강조한다.

〈표 2-7〉 국내 주요 대학의 생성형 인공지능 활용 가이드라인

대학명	주요내용
고려대학교	ChatGPT 등 AI의 기본 활용 가이드라인(23. 3. 15) • 목적: 학습자의 학습 효과 및 긍정적 교육 경험을 위해 생성형 인공지능을 활용 권리 보장, 가이드라인 부재 시 교수·학습에 부정적 영향 우려 • 대표 사례: 표절, 부정행위와 같은 학문적 진실성(academic integrity) 위반, 비판적·창의적 사고 및 문제해결 능력 저하, 부정확하고 편향된 정보 습득, 고정관념의 강화 • 권장 사항: 인공지능의 윤리적 사용을 강조하는 교육 실시, 학생의 비판적 사고 능력 개발, 과제 수행 시 전반적인 과정에 대해 상세한 기재 등
연세대학교	생성형 AI 활용 가이드라인(24. 5) • 긍정적 이용을 위한 요건: ① 비판적 활용, ② 상호합의적 활용, ③ 창의적 활용, ④ 윤리적 활용, ⑤ 보안적 활용, ⑥ 개방적 활용 • 교수자, 학습자, 그리고 연구자를 위한 세부 활용 방안 지침 제공 • 수업 및 과제 내 생성형 인공지능 허용 여부 확인, 명확한 출처 표기, 정보의 진위성 확인 등을 권장 • 생성형 인공지능 활용 시 유의해야 할 사항: 개인정보와 저작권, 보안 문제 등
성균관대학교	챗GPT 등 AI와 학습윤리(2023. 3. 10) • 챗GPT 사용에서 발생가능한 문제들: 표절, 팩트체크 부재, 편향성 강화 등 • 챗GPT와 학습윤리: ① 챗GPT의 결과물을 자신이 직접 한 과제인양 제출하는 것은 부정행위, ② 챗봇의 답변에 의존하지 않고 정확한 정보를 탐색하며 윤리적 책임을 지닐 것 • 윤리적 사용을 위한 능력: '질문'하는 능력, 챗GPT의 글에 대해 '비판'하는 능력 챗GPT 종합안내 홈페이지(24. 6. 10, 접속). • 챗GPT 종합안내 홈페이지 개설 목적: 인공지능의 교육적 선용(善用), 교육 모델 개발, 부정 행위·표절 적극 예방, 인공지능 활용 윤리·디지털 리터러시 교육 강화, 적절한 평가 방식 개발 등 챗GPT의 교육적 활용(홈페이지 소개, 24. 12. 11, 접속) • 챗GPT 기본 정보, 최신 활용 사례들, 사용 방법 등을 안내 • 다양한 부정행위 사례와 교강사 대응 가이드(평가 방법, 시험 환경 통제 및 관리, 인공지능 탐지 프로그램 활용), 올바른 활용과 학습윤리 등을 제공

대학명	주요내용
중앙대학교	생성형 AI 활용 가이드라인(24. 4. 9. 개정) • 가이드라인 목적: 생성형 인공지능 확산에 대한 대응, 생성형 인공지능 활용 시 교수 및 학습 윤리 준수 • 주제 별 지침 구분: ① 교수자 및 학습자 공통, ② 교수자, ③ 학습자 • 상호 간 합의 후 사용, 생성형 인공지능 기술에 대한 이해, 정보의 진위 확인, 표절과 부정 사용 금지 등 연구 윤리 준수 등이 공통으로 권장됨
UNIST	생성형 AI 활용 가이드라인(23.07.28) • 가이드라인 구성: ① 생성형 인공지능 기술에 대한 소개, ② 교원, 학생, 연구원을 위한 생성형 인공지능 활용 가이드라인, ③ 생성형 인공지능 활용 팁 • 교수용 가이드라인: ① 교육: 직접 사용해보기, 활용 여부는 수업의 특성을 고려, 허용 여부를 사전 계획서에 명시, 다양한 형태의 평가 방식 활용, 인용 및 출처 표기 요구, 필요 시 인공지능 탐지기 사용, ② 연구: 학술지/학회의 생성형 인공지능 사용 가이드라인 확인, 생성형 인공지능 인용, 출처 확인, 편향성 주의, 사실 여부 확인, ③ 보안: 보안이 필요한 중요연구는 인공지능에 공유 금지, 최신 보안 소프트웨어 업데이트 확인, 개인정보 공유 금지, 생성된 코드 사용 시 주의 • 학생용 가이드라인: ① 학습: 수업 정책 준수, 사실 여부 확인, 비판적 사고, 보조도구로써만 활용, ② 연구: 학술지/학회 가이드라인 확인, 사용 시 인용, 출처 확인, 편향성 주의, ③ 보안: 보안 필요 중요 연구 정보는 인공지능에 공유 금지, 보안 소프트웨어 최신화, 개인정보 공유 금지, 생성된 코드 사용 주의 • 연구원용 가이드라인: ① 연구: 학술지/학회 가이드라인 확인, 사용 시 인용, 출처 확인, 편향성 주의, ② 보안: 보안 필요 중요 연구 정보는 인공지능에 공유 금지, 보안 소프트웨어 최신화, 개인정보 공유 금지, 생성된 코드 사용 주의

자료: 연구진 작성

각 대학의 생성형 인공지능 이용 가이드라인과 대처에 대한 주요 내용 요약은 아래와 같다.

－ 고려대학교(ChatGPT 등 AI의 기본 활용 가이드라인, 23. 3. 15.): 학습자의 학습 효과 및 긍정적 교육 경험을 위해 생성형 인공지능을 활용할 수 있는 권리 보장을 명시하고 있다. 하지만, 기본적인 가이드라인 없이 수업에 생성형 인공지능을 사용할 시, 교수·학습의 질에 부정적 영향을 미칠 수 있으며, 대표적인 사례로 표절, 부정행위와 같은 학문적 진실성(academic integrity) 위반, 비판적·창의적 사고 및 문제해결 능력 저하, 부정확하고 편향

된 정보 습득, 고정관념의 강화 등을 제시하고 있다. 인공지능의 윤리적 사용을 강조하는 교육 실시, 학생의 비판적 사고 능력 개발, 경험적 데이터 수집 및 분석이 필요한 과제 제시, 과제 수행 시 전반적인 과정을 상세히 기재 등 세부 지침을 제시하고 있다.

－ 연세대학교(생성형 AI 활용 가이드라인, 24.05.): 생성형 인공지능으로 인한 교육 및 연구 분야의 변화를 설명하며 긍정의 효과로, ① 자료 수집 및 콘텐츠 생성, ② 다양한 언어적 활용, ③ 창조적인 일 수행, ④ 창의적 아이디어 도출을 예시로 하고 있다. 교육 및 연구 분야에서 생성형 인공지능을 긍정적으로 이용할 수 있도록 ① 비판적 활용, ② 상호합의적 활용, ③ 창의적 활용, ④ 윤리적 활용, ⑤ 보안적 활용, ⑥ 개방적 활용을 명시하고 있다. 교수자, 학습자, 그리고 연구자를 위한 세부 활용 방안 지침을 제시하고 있으며 수업 및 과제 내 생성형 인공지능 허용 여부 확인, 명확한 출처 표기, 정보의 진위성 확인 등을 권장하고 있다. 개인정보와 저작권, 보안의 문제 또한 생성형 인공지능 활용 시 유의해야 할 사안으로 안내하고 있다.

－ 성균관대학교(챗GPT 종합안내 홈페이지, 24. 6. 10., 접속): 성균관대학교는 챗GPT 종합안내 홈페이지를 개설하여 인공지능의 교육적 선용(善用), 교육 모델 개발, 부정 행위·표절 적극 예방, 인공지능 활용 윤리·디지털 리터러시 교육 강화, 적절한 평가 방식 개발 등 챗GPT의 교육적 활용을 위한 노력을 시행 중이다. 챗GPT 기본 정보, 최신 활용 사례들, 사용 방법 등을 안내하고 있으며, 다양한 부정행위 사례와 교강사 대응 가이드(평가 방법, 시험 환경 통제 및 관리, 인공지능 탐지 프로그램 활용), 올바른 활용과 학습윤리 등을 제공하고 있다.

－ 성균관대학교(챗GPT 등 AI와 학습윤리, 2023. 3. 10.): 챗GPT 사용으로 인해 발생 가능한 문제들로, 표절, 팩트체크 부재, 편향성 강화를 설명하고 있다. 챗GPT와 학습윤리로서, ① 챗GPT의 결과물을 자신이 직접 한 과제 인양 제출하는 것은 부정행위, ② 챗봇의 답변에 의존하지 않고 정확한 정보를 탐색하며 윤리적 책임을 지닐 것을 명시하고 있다. ‘질문’하는 능력, 챗GPT의 글에 대해 ‘비판’하는 능력을 갖출 것을 강조한다.

－ 중앙대학교(생성형 AI 활용 가이드라인, 24. 4. 9. 개정): 생성형 인공지능 확산에 대한 대응으로, 생성형 인공지능 활용 시 교수 및 학습 윤리를 준수할 수 있도록 가이드라인을 제시하고 있다. 생성형 인공지능의 정의와 종류, 긍·부정적 측면을 설명하며, ① 교수자 및 학습자 공통, ② 교수자, ③ 학습자로 구분하여 지침을 제시하고 있다. 해당 지침들은

상호 간 합의 후 사용, 생성형 인공지능 기술에 대한 이해, 정보의 진위 확인, 표절과 부정 사용 금지 등 연구 윤리 준수를 내포하고 있다.

－ UNIST(생성형 AI 활용 가이드라인, 23. 7. 28): 생성형 인공지능의 등장이 학내 구성원들의 교육, 학습, 연구 활동에 미치는 영향을 고려하여 설문조사 실시 후 활용방안을 가이드북으로 제시하였다. 가이드라인은 크게 세 부분으로 구성되며, ① 생성형 인공지능 기술에 대한 소개, ② 교원, 학생, 연구원을 위한 생성형 인공지능 활용 가이드라인, ③ 생성형 인공지능 활용 팀이 그 내용이다. 부록으로, UNIST 교원, 학생, 연구원을 대상으로 생성형 인공지능 사용실태 및 인식조사 결과가 담겨있다. 2023년 4월 28일부터 5월 7일까지, UNIST 교내 구성원을 대상으로 시행한 ‘생성형 AI 사용실태 및 인식조사’의 주요 결과로, 대학원생의 93%, 학부생의 90%는 ChatGPT 등 생성형 인공지능을 사용해 본 적이 있으며, 학부생의 65%는 수업에 생성형 인공지능을 활용하고 있다. 교수의 경우 “향후 수업에 생성형 인공지능을 활용할 의향이 있다”란 질문에 그렇다(52%), 아니다(14%)로 응답하였다. 또한, 교수들은 “과제/시험에 생성형 인공지능 사용을 금지할 예정이다”에 그렇다(18%), 아니다(65%)의 응답을 하였지만, 동시에 “과제/시험 방식의 변경이 필요하다”에 그렇다(56%), 아니다(26%)를 보임으로써 생성형 인공지능 사용을 금지하진 않지만, 학생들의 성적을 평가하는 방식에 대한 새로운 변화가 필요함을 인지하고 있다.

나. 해외 주요 대학 생성형 인공지능 활용 가이드라인

해외의 대학 역시 생성형 인공지능 활용에 대한 긍정적 효과에 주목하고 그에 대한 가이드라인을 제공한다. 국내 대학의 경우와 비교할 때 특히 주목되는 특징은 해외 사례의 경우 학생들이 안전한 생성형 인공지능 사용을 지원하는 담당 부서와 문의처뿐만 아니라 시중에 나와 있는 여러 서비스를 비교하는 정보 제공 부분이다. 예를 들어, 하버드 대학의 인공지능 관련 웹페이지는 여러 인공지능 모델 및 서비스의 특징을 비교하여 적합한 모델을 선택하도록 권장하고 있다. MIT는 학교 차원에서 생성형 인공지능 서비스의 안전한 이용을 장려하기 위해 생성형 인공지능 모델의 라이선스를 구매하고 있다. 이러한 사례들을 통해 해외의 주요 대학에서는 활용 지침만이 아니라 실질적인 지원도 실행되고 있음을 알 수 있다. 또, 윤리적이고 책임감 있는 사용을 안내하는데, 특히 ‘학술적 진실성’에 대해 공통으로 강조하고 있고, ‘정보보안’과 ‘민감 데이터’도 핵심 이슈로 다루고 있다.

〈표 2-8〉 해외 주요 대학의 인공지능 활용 가이드라인 사례

대학명	주요내용
옥스퍼드 대학	학습 지원을 위한 생성형 인공지능 도구 사용법(Use of generative AI tools to support learning, 24. 12. 11. 최종 접속) • 생성형 인공지능의 결과물을 비판적으로 검토함으로써 유용한 학습 보조 수단으로 활용할 것을 권장
하버드 대학	생성형 인공지능 도구 이용에 대한 초기 가이드라인(Initial guidelines for the use of Generative AI tools at Harvard, 24. 12. 11. 최종 접속) • 책임감 있는 생성형 인공지능 활용 권장 • 생성형 인공지능 이용 중 정보 보안, 사생활 데이터 문제, 저작권, 규정, 학술성 진실성 지향 • 교내 관련 부서를 통해 기밀 데이터 보호, 기재 전 검토, 학문적 정책 준수, 피싱사기 주의, 생성형 인공지능 툴 지원 등을 안내 • 인공지능 활용 사례 및 인공지능 툴 간 비교, 유용한 텍스트/이미지 기반 인공지능 프롬프트와 같은 효과적인 인공지능 활용을 위한 방안 제시 • 생성형 인공지능(Generative Artificial Intelligence, 24. 12. 11. 최종 접속) • 인공지능 툴 간 비교, 텍스트/이미지 기반 인공지능에 효과적인 프롬프트와 같이 효과적인 인공지능 활용을 위한 방안을 제시 • 학교 홈페이지에서 공식적으로, 사용 가능한 인공지능 툴과 각각에 대한 설명, 이용 가능 대상자, 데이터 분류 단계 등을 제공
MIT 공대	생성형 인공지능 도구 사용에 대한 최초 안내(Initial guidance for use of Generative AI tools, 24. 4. 3. 개정) • 정보 보안, 사생활 데이터, 규제와 정책 준수, 지적 재산권, 학술적 진실성을 고려해 지침 제시 • 교내 정보/보안 데이터 보호, 투명한 사용, 정보의 정확성 확보 • 담당 부서를 통한 생성형 인공지능 사용 지원
스탠포드 대학	생성형 인공지능 정책 안내(Generative AI Policy Guidance, 23. 2. 16. 채택) • 생성형 인공지능의 명과 암에 대한 경계 • 생성형 인공지능 사용 시, 사전에 교수자의 동의를 구할 것 • 교수자는 생성형 인공지능 사용에 대한 기준을 명확히 제시
예일 대학	생성형 인공지능 도구 사용에 대한 가이드라인(Guidelines for the Use of Generative AI Tools, 23. 9. 20. 게재) • 생성형 인공지능 사용으로 인한 가능성과 위험성에 대한 인지 • 안전성, 보안성, 학술적 진실성 준수 • 학교와 개인의 민감한 정보 및 데이터 주의 • 학술적 진실성을 준수, 편견과 부정확성 경계 • 담당 부서를 통한 생성형 인공지능 사용 지원

자료: 연구진 작성

해외 주요 대학의 생성형 인공지능 가이드라인 주요 내용은 아래와 같다.

－ 옥스퍼드 대학(학습 지원을 위한 생성형 인공지능 도구 사용법, Use of generative AI tools to support learning, 24. 12. 11. 최종 접속): 인공지능 도구들은 인간의 비판적 사고와 대학 교육의 근간을 이루는 학문적 증거 기반 논증 및 지식 개발을 대체할 수 없음을 알고 있다. 학생들은 학문적 능력을 발전시키기 위해 생성형 인공지능 툴을 이용할 수 있지만, 생성형 인공지능의 결과물을 비판적으로 검토함으로써 학습에 유용한 보조적 수단으로 활용해야 한다고 명시하고 있다.

－ 하버드 대학(생성형 인공지능 도구 이용에 대한 초기 가이드라인, Initial guidelines for the use of Generative AI tools at Harvard, 24. 12. 11. 최종 접속): 생성형 인공지능 툴을 이용하는 데 있어 책임감 있는 사용을 지원하고 있다. 생성형 인공지능 툴을 이용하는 데 있어 정보 보안, 사생활 데이터 문제, 저작권, 규정, 그리고 학술적 진실성 등을 고려하는 것이 중요함을 밝히고 있다. 기밀 데이터 보호, 기재 전 검토, 학문적 정책 준수, 피싱 사기 주의, 생성형 인공지능 툴 구매 전 교내 문의와 같은 상세한 지침이 있으며, 교내 관련 부서 및 연락처를 안내하고 있다.

－ 하버드 대학(생성형 인공지능, Generative Artificial Intelligence, 24. 12. 11. 최종 접속): 인공지능 활용 사례 및 인공지능 툴 간 비교, 텍스트/이미지 기반 인공지능 툴에 효과적인 프롬프트와 같은 효과적인 인공지능 활용을 위한 방안을 제시하고 있다.

－ MIT 공대(생성형 인공지능 도구 사용에 대한 최초 안내, Initial guidance for use of Generative AI tools, 24. 4. 3. 개정): 교내 구성원이 정보 보안, 사생활 데이터, 규제와 정책 준수, 지적 재산권, 학술적 진실성을 고려하여 생성형 인공지능을 사용할 수 있도록 가이드의 목적을 밝히고 있다. 교내의 정보와 보안 데이터를 보호하고, 인공지능 툴 사용에 대해 투명하게 밝히며, 인공지능으로 생성한 정보에 대해 정확성을 확보하는 것을 주된 지침으로 제공하고 있다. 생성형 인공지능 툴을 구매하거나 사용하기 전 관련 부서에 연락하여 지원 받을 수 있음을 명시하고 있다.

－ 스탠포드 대학(생성형 인공지능 정책 안내, Generative AI Policy Guidance, 23. 2. 16. 채택): 생성형 인공지능 툴의 사용은 학생의 학습과 이해를 증진할 수 있지만, 주요 학습 목표인 인간 사고의 증진을 성취하지 못할 수도 있음을 경계하고 있다. 교수자와 협의 없

이 혹은 확실한 명시가 이루어지지 않은 채 과제나 시험에 생성형 인공지능을 사용하지 않아야 함을 알리고 있다. 교수자는 자신의 수업에서 생성형 인공지능 사용에 대한 기준을 자유롭게 제시할 수 있으며, 강의 안내서에 명시하거나 학생들에게 명확히 전달해야 한다고 기재하였다.

－ 예일 대학(생성형 인공지능 도구 사용에 대한 가이드라인, Guidelines for the Use of Generative AI Tools, 23. 9. 20. 게재): 생성형 인공지능의 사용으로 인한 긍정적 가능성과 부정적 위험을 동시에 언급하며, 안전성, 보안성, 학술적 진실성을 위한 가이드라인 준수를 권고하고 있다. 학교와 개인에게 민감한 정보 및 데이터 사용에 대해 주의하고 있다. 학술적 진실성을 준수하며 편견과 부정확성에 대해 경계하여야 한다고 강조한다. 인공지능 사용에 대해 지원을 받거나 위험을 피하기 위해 담당자에게 연락하여 도움받을 수 있도록 조치하고 있다.

제2절 생성형 인공지능 서비스 규제 동향

이번 절에서는 국내의 인공지능 서비스 규제 동향을 살펴본다. 인공지능 서비스는 기술의 불확실성과 불완전성으로 인해 우선 경성법보다는 연성법으로 관리해야 한다는 사회적 합의가 이루어졌고, 이에 지난 2017년 이후 국제기구 및 각국, 그리고 사업자별로 인공지능 윤리를 개발하여 제시해왔다. 2022년 생성형 인공지능이 일반 서비스로 선보인 이후에는 인공지능을 법으로 규제해야 한다는 목소리가 더 높아졌으며, 이에 2024년 8월 유럽연합(EU)의 인공지능법(AI Act)이 발효되었고, 우리나라를 포함하여 캐나다와 미국 등 여러 국가에서 인공지능 관련 법안을 발의해왔다. 여기서는 주목할 만한 주요 국제기구 및 국가들의 규제 동향을 살펴본다.

1. 국제기구

먼저 국제기구의 규제 동향이다. UN은 2024년 3월 21일 ‘지속가능한 발전을 위한 안전하고 신뢰할 수 있는 인공지능 시스템의 기회 포착(Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development)’에 관한

결의안을 채택하였다. 2024년 9월경에는 유럽 평의회(Council of Europe)가 주도하는 유럽 이사회 기본협약(Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law)에 유럽연합 회원국, 미국, 일본 등이 참여한 바 있다. OECD, G7 및 인공지능 안전 정상회의 등의 국제기구도 인공지능 관련 해석례와 프레임워크 등을 제안하고 있다.

다만, 국제기구는 제재 수단이 결부된 결의안이나 협약을 제정하고 있지는 않으며, AIAB(Artificial Intelligence Advisory Body), UNESCO 등 산하 전문기관을 통해 주로 인공지능 관련 가이드라인 또는 연구보고서를 발표하고 있다.

가. UN

- 1) 지속가능한 발전을 위한 안전하고 신뢰할 수 있는 인공지능 시스템의 기회 포착 (Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development, 2024. 3. 21.)

본 결의안은 UN 총회가 인공지능 분야 규제를 위해 채택한 최초의 결의안으로서 우리나라를 비롯하여 120여 개 회원국의 지지를 받았다. 본 결의안은 '비군사 영역에 속하며, 사전 설계, 설계, 개발, 평가, 테스트, 배치, 사용, 판매, 조달, 운영, 폐기의 생애 주기 단계가 포함된 인공지능 시스템'에 적용된다.

본 결의안은 개발도상국에 대한 지원과 국제 협력을 장려하고 인권법 준수를 촉구하며, 위험 관리, 데이터 보호, 투명성, 인공지능 시스템의 언어적·문화적 다양성 보존을 위한 조치 사항도 명시되어 있다. 특히 콘텐츠 조작에 대한 위험 관리 측면에서, 인공지능 생성 콘텐츠의 식별을 위한 워터마킹이나 라벨링 등 '콘텐츠 인증 및 출처 확인 매커니즘'의 개발 필요성이 명시되어 있다.

- 2) UNESCO, 인공지능 윤리 권고(Recommendation on the Ethics of Artificial Intelligence, 2021. 11. 23.)

UNESCO는 인공지능과 관련한 긍정적 부정적 영향을 파악하고 인간 존엄성, 인권, 지속가능한 발전, 포용성 등을 중심으로 윤리적 프레임워크를 제시하고 있다. 인공지능 시스템의 윤리적인 문제는 연구, 설계, 개발, 배포, 사용까지의 전체 생애 주기와 관련된 것이다.

본 권고안은 위와 같은 가치와 원칙을 구현하기 위한 여러 정책영역 발굴하고 관련 행

동을 촉구하고 있는데, 이 중 ‘커뮤니케이션 및 정보’ 정책 영역에서는 허위정보, 오정보, 혐오 발언을 완화하고 대응하기 위한 디지털 및 미디어 정보 리터러시 기술에 대한 투자와 증진을 강조하고 있다.

3) AIAB, 인류를 위한 인공지능 거버넌스(Governing AI for Humanity, 2024. 9. 19.)

본 보고서는 글로벌 인공지능 거버넌스 격차를 우려하고 있으며, 이에 대한 해결책으로 인공지능 관련 국제적인 과학 전문가 집단 모임, 인공지능 거버넌스에 관한 정책적 대화 활성화, 인공지능 거버넌스 표준 개발, 인공지능 역량 개발에 관한 글로벌 네트워크 구축, 인공지능을 위한 글로벌 기금 마련, 글로벌 인공지능 데이터 프레임워크 개발, UN 내 인공지능 거버넌스 전담 사무국 설치 등 7가지 권고 사항을 제시한다. 특히 글로벌 인공지능 데이터 프레임워크 개발과 관련하여 인공지능 학습 데이터의 문화적·언어적 다양성 증진과 인공지능 학습 데이터의 출처 및 사용에 관한 공통 표준의 수립을 강조하고 있다.

나. G7의 히로시마 프로세스(Hiroshima Process, 2023. 10. 30.)

G7 정상들은 첨단 인공지능 시스템을 개발하는 조직을 위한 히로시마 프로세스 지도 원칙(guiding principles)과 이러한 원칙에 기반한 행동 강령(code of conduct)에 합의하였다. 위 행동강령은 첨단 인공지능 시스템(advanced AI systems) 개발자에게 주로 적용된다.

첨단 인공지능 시스템 개발자에게는 ① 인공지능 시스템 생애 주기 전반에 걸친 위험 식별·평가·완화 조치, ② 인공지능 시스템 배포 후 취약성·사고·오용 패턴 식별 및 완화 조치, ③ 첨단 인공지능 시스템의 기능·한계·적절/부적절 사용영역 공개 조치, ④ 인공지능 시스템 사용 조직 간 책임감 있는 정보 공유 및 사고 보고, ⑤ 위험 기반 접근 방식에 기반한 인공지능 거버넌스 및 위험 관리 정책 개발, ⑥ 강력한 보안 통제, ⑦ 신뢰할 수 있는 콘텐츠 인증 및 출처 증명 메커니즘을 개발 및 배포, ⑧ 사회·안전·보안 위협의 완화를 위한 연구 및 투자 우선순위 설정, ⑨ 국제 문제 해결을 위한 첨단 인공지능 시스템 개발, ⑩ 국제 기술 표준의 개발과 채택, ⑪ 적절한 데이터 입력을 통한 개인 정보 및 저작권 보호 조치 등 11가지의 조치 사항을 권고하고 있다. 다만 행동 강령에서는 ‘각 관할권마다 이러한 조치를 시행하는 데 있어 각기 다른 방식으로 고유한 접근 방식을 취할 수 있다’고도 명시하고 있다.

③항의 투명성 항목과 관련하여, 새로운 첨단 인공지능 시스템을 출시할 때마다 ‘투명성

보고서'를 발행하여 시스템에 대한 중요 정보를 제공해야 하며, 중요 정보에는 '인권 등을 침해하는 위험에 대한 평가', '성능상의 중요한 제한 사항', '편향·차별 등 인공지능 시스템이 사회에 미칠 수 있는 잠재적 영향과 위험에 대한 평가', '인공지능 시스템 적합성 평가에 관한 레드팀 활동 결과'가 포함된다.

⑦항의 콘텐츠 인증과 관련하여, 출처 데이터에는 콘텐츠를 만든 서비스 또는 모델의 식별자가 포함되어야 하나 사용자 정보는 포함되지 않아야 한다. 라벨링과 면책 약관(disclaimer)을 통해 사용자가 인공지능 시스템과 상호작용하고 있음을 알리도록 하는 것도 필요하다.

다. OECD, 인공지능 권고안(Recommendation of the Council on Artificial Intelligence, 2024. 5. 5. revised)

OECD는 2019년 5월 22일에 채택한 OECD AI 위원회의 권고안을 2024년 5월 5일부로 개정하였는데, 권고안은 인공지능 시스템의 책임감 있는 개발 및 활용을 위한 OECD의 정책 권고 사항을 담고 있다.

개정 권고안은 인공지능 시스템을 '명시적 또는 암시적 목표를 위해 입력값을 받아 물리적 또는 가상 환경에 영향을 줄 수 있는 예측, 콘텐츠, 추천 또는 결정과 같은 출력을 생성하는 방법을 추론하는 기계 기반 시스템'이라고 정의하고 있으며, 인공지능 시스템은 배포 후 자율성 및 적응성 수준이 다르다고 설명하고 있다. '생성형 인공지능' 개념을 포섭하기 위해 '콘텐츠' 출력을 인공지능 시스템 정의에 명시적으로 포함한 것이다.

본 권고안은 인공지능 시스템의 투명성과 설명가능성을 증진하여 책임성을 확보할 것을 주요 목표로 하고 있는데, 개정 핵심 내용은 ① 생성형 인공지능으로 인해 심화되는 허위 정보와 정보 조작 문제에 대응하기 위한 무결성 보호(표현의 자유 존중 보장 필요) 및 책임 있는 정보 공개 조치, ② 인공지능 시스템의 오용 방지를 위한 위험 완화 조치, 인공지능 시스템이 과도한 피해 등을 입히는 경우에 인간의 개입을 통해 안전하게 시스템을 재정의·수리 또는 해체할 수 있는 안전 메커니즘 등을 구축함으로써 인공지능 시스템의 책임성을 강화한 것이다.

라. 인공지능, 민주주의, 법치주의를 위한 협약(Convention on Artificial Intelligence and Human Democracy and the Rule of Law, 2024. 9.)

유럽 평의회의 인공지능위원회(CAD)는 인권, 민주주의, 법치주의에 초점을 맞추어 인공지능 관련 기본협약을 마련하였고, 위 기본협약에 영국, EU 회원국, 미국 등이 서명하였다. 본 기본협약은 인공지능에 관한 최초의 법적 구속력이 있는 국제 협약이다. 다만, 협약 당사국으로 하여금 협약에 따른 준수 및 조치에 대해 보고하도록 하고 협약 당사국이 자체적으로 협약 준수 여부를 감독하기 위한 독립 기관을 지정하도록 하고 있으나, 협약 위반에 따른 제재 사항을 규정하고 있지는 않다.

본 기본협약은 국가 안보 이익과 관련된 인공지능 시스템, 연구 및 개발 중인 인공지능 시스템, 국방과 관련된 인공지능 시스템을 적용 범위에서 제외하고 있으며, 공공 부문의 인공지능 시스템의 생애 주기에 대해서는 기본협약상의 원칙과 의무가 적용된다고 하고 있으나, 민간 부문에 대해서는 각국이 협약상의 원칙과 의무를 적용할 것인지 아니면 다른 적절한 조치를 취할 것인지 결정할 수 있도록 하고 있다.

기본협약 제8조(투명성 및 감독)에 의하면, 각 협약 당사자는 인공지능 시스템에 의해 생성된 콘텐츠의 식별을 포함하여 인공지능 시스템의 생애 주기 내 활동과 관련하여 특정 상황 및 위험에 맞는 적절한 투명성 및 감독 요건을 보장하기 위한 조치를 채택하거나 유지하여야 한다. 즉, 인공지능 시스템이 사용되는 맥락과 위험성을 고려하여 투명성과 감독 수준을 결정하는 유연한 접근 방식을 취하고 있는 것이다.

또한 기본협약 제10조(평등 및 차별 금지)에 의하면, 각 당사국은 성평등 및 차별 금지를 포함한 평등을 존중하고 보장하는 조치를 채택하거나 유지해야 하며 불평등을 극복하여 공평하고 공정하며 평등한 결과를 도출하는 조치를 채택하거나 유지하여야 한다.

기본협약 제16조(위험 및 영향 관리 프레임워크)에서는, 협약 당사국이 인공지능 시스템이 야기하는 '실재하거나 잠재적인 위험'을 식별·평가·예방 및 완화하는 조치를 취해야 한다고 규정한다. 특히 인공지능 시스템이 인권, 민주주의, 법치주의에 미치는 영향을 고려해야 하며, 이러한 위험 관리 조치는 필요에 따라 점진적이고 차별적으로 적용되어야 한다고 하면서, 위험, 실제적 및 잠재적 영향, 위험 관리 방식에 대한 문서화를 해야 하며, 인공지능 시스템을 크게 수정하는 경우에는 테스트를 진행해야 한다.

1) 인권, 민주주의, 법치주의를 위한 방법론 (HUDERIA methodology, 2024. 11. 28.)

유럽 인공지능 위원회는 본 방법론을 통해 인공지능 시스템의 위험 및 영향 평가에 대한 구조화된 접근 방식을 제공하며, 특히 인권·민주주의·법치주의 보호 및 증진에 초점을 맞춘다. 본 방법론은 비록 법적 구속력은 없으나, 위 EU 회원국 등 기본협약 당사국이 위험 평가에 대한 접근 방식 개발 시 지침으로 활용할 수 있다.

본 방법론은 네 가지 요소로 구성되어 있고, ① 맥락 기반 위험 분석(COBRA)은 인공지능 시스템이 적절한 솔루션인지에 대한 초기 판단을 포함하여 인공지능 시스템의 위험을 식별하고 이해하는 데 필요한 정보를 수집하고 매핑하는 구조화된 접근 방식을 제공하며, ② 이해관계자 참여 프로세스(SEP)는 잠재적으로 영향을 받을 수 있는 사람들에 관한 정보를 얻고, 잠재적 피해 및 완화 조치를 맥락화하거나 확증하기 위해 관련 이해관계자와의 참여를 적절히 활성화하고 운영하는 접근 방식을 제안하며, ③ 위험 및 영향 평가(RIA)는 인공지능 시스템의 생애 주기 내 활동이 인권, 민주주의, 법치주의에 미칠 수 있는 잠재적 및 실제적 영향에 대한 자세한 평가 방식을 제공하고, ④ 완화 계획(MP)은 구체 수단 제시 및 반복 검토를 포함하여 완화 및 시정 조치를 정하는 방식을 제공한다.

마. AI SUMMIT

1) 블레츨리 선언(Bletchley Declaration, 2023. 11. 1.)

우리나라, 미국, EU, 중국을 포함한 28개 관할권이 참여한 제1회 ‘글로벌 AI 안전 서밋’은 ‘AI 안전에 관한 블레츨리 선언’을 채택하였다. 본 선언문은(고도의 범용 인공지능 모델과 해를 입힐 수 있는 기능이 포함된 인공지능으로 이해되는) ‘프론티어 AI’가 야기할 수 있는 심각한 안전 위험을 우려하고 있으며, 이러한 위험에 대처하기 위해 위험 기반 정책 구축, 민간 부문의 투명성 강화, 적절한 평가 지표 및 안전성 테스트 도구 개발 등이 필요하다고 강조한다.

2) 서울 선언(Seoul Declaration, Seoul Statement of Intent, 2024. 5. 21.)

본 선언문은 안전하고 혁신적이며 포용적인 인공지능 생태계를 조성하는 위험 기반 접근법을 포함한 정책 및 거버넌스 프레임워크를 지지한다. 프레임워크는 인간의 창의성과 인공지능의 개발 및 사용 간의 선순환을 촉진하고, 사회 문화적·언어적·성별 다양성을 증진하며, 상업적 및 공개적으로 이용 가능한 인공지능 시스템의 생애 주기 동안 환경적으

로 지속 가능한 기술 및 인프라의 개발과 사용을 촉진해야 한다.

본 선언문의 부속서로 채택된 의향서에 의하면, 인공지능 안전 연구 및 테스트, 인공지능 안전 지침 개발을 수행하는 인공지능 안전연구소를 설립하고 지원하는 것이 필요하며, 국제적인 인공지능 안전 보고서(AI safety report)를 활용하여 참여국의 거버넌스 프레임워크 수립 및 조정을 지원할 예정이다.

3) 첨단 인공지능의 안전에 관한 국제 과학 보고서(International Scientific Report on the Safety of Advanced AI - Interim Report, 2024. 5. 17.)

영국 과학혁신기술부(DSIT)는 인공지능 정상회의에 참여한 국가 및 국제기구의 패널과 협력하여 ‘첨단 AI의 안전에 관한 국제 과학 보고서(중간 보고서)’를 발표했다. 이 보고서는 범용 인공지능의 잠재적 위험을 분석하고 이러한 위험을 완화하기 위해 사용할 수 있는 기술적 방법에 대해 논의하고 있다.

본 보고서는 첨단 인공지능과 관련된 위험을 악의적인 사용으로 인한 위험(허위 콘텐츠 및 사기, 허위 정보 및 여론 조작 등), 오작동으로 인한 위험(데이터 편향 및 환각 현상 등), 시스템적 위험(환경 및 개인정보 보호 문제 등)으로 구분하고 있으며, 이러한 위험의 완화 방안으로 ① 위험 관리 및 안전 엔지니어링(위험 평가 및 관리 프로세스), ② 신뢰할 수 있는 모델 학습(적대적 훈련, 환각과 같은 유해한 행동 방지 노력), ③ 모니터링 및 인간 개입(이상 징후 탐지 모니터링), ④ 공정성 및 편견 완화(훈련 데이터 다양성, 공정성 지표 및 투명성 개선), ⑤ 개인정보 보호(데이터 최소화 등) 조치를 권고한다.

2. 유럽연합(EU)

EU의 인공지능 거버넌스 체계는 광범위한 디지털 거버넌스 규제 프레임워크와 밀접하게 상호작용을 하고 있다. 특히 EU의 인공지능법(AI ACT)을 중심으로 개인정보(GDPR), 저작권(Copyright Directive), 디지털 시장 법안(Digital Market Act), 디지털 서비스 법안(Digital Services Act; 온라인 플랫폼 등), 제조물 책임(Product Liability Directive), 데이터 거버넌스 및 데이터 공유 법안(Data Governance Act, Data Act), 전산망과 정보 시스템(Network and Information Systems 2 Directive), 디지털/사이버 탄력성 법안(the Digital Operational Resilience Act, Cyber Resilience Act; CRA, Critical Entities Resilience(CER)

Directive), 사이버 보안 법안(Cybersecurity Act), 플랫폼 노동자(EU Platform Workers Directive) 등의 기존 규정과 연관성이 깊다. 이하에서는 EU AI ACT를 중심으로 인공지능 규제를 살펴본다.

가. EU의 인공지능법(AI ACT)

EU의 인공지능법(AI ACT)는 모든 EU 회원국에 직접 적용되는 규정으로서, 인공지능에 대한 규칙을 보장하고 조화시키는 것을 목표로 한다. EU의 일반개인정보보호법(GDPR)이 개인의 프라이버시 및 데이터 보호 권리를 보호하는 데 중점을 둔 반면, EU 인공지능법의 주요 목적은 인공지능을 규율하는 조화롭고 일관된 법적 프레임워크를 확립하고 기본권과 민주주의 존중 등 EU의 가치에 부합하는 방식으로 인공지능 기술을 안전하게 개발 및 사용하도록 보장하는 것이다.

EU 인공지능법 제3조 제1항은 인공지능 시스템을 '다양한 수준의 자율성을 가지고 작동하도록 설계된 기계 기반 시스템으로, 배포 후 적응력을 발휘할 수 있으며 명시적 또는 암묵적 목적을 위해 수신한 입력으로부터 물리적 또는 가상 환경에 영향을 미칠 수 있는 예측, 콘텐츠, 추천 또는 결정과 같은 출력을 생성하는 방법을 추론하는 시스템'이라고 정의하여, 전술한 OECD의 인공지능 시스템 정의에 부합하게 규정되어 있다. 위 정의에 '콘텐츠' 생성이 포함되므로, 인공지능 시스템 정의에는 생성형 인공지능 시스템이 포함된다.

EU 인공지능법은 네 가지 위험 범주로 인공지능 시스템을 분류하고 각 범주에 따라 인공지능 생애 주기의 관련 주체가 취해야 할 다양한 위험 기반 조치를 규정한다. 입법 과정에서 범용 인공지능에 대한 요구 사항이 추가되어 사실상 다섯 가지의 범주로 구분되었지만, 위 분류는 서로 중첩될 수 있다. EU 인공지능법은 전문(recital) 제99항에서 대규모 생성형 인공지능 모델(large generative AI model)이 텍스트, 오디오, 이미지, 동영상 등 다양한 형태의 콘텐츠를 유연하게 생성할 수 있어 넓은 범위의 특정 작업에 쉽게 적용할 수 있다는 점에서 '범용 인공지능'의 전형적인 예시라고 한다.

범용 인공지능 모델이 다양한 인공지능 시스템에 통합되어 영향을 미칠 수 있다는 점을 고려하여, EU 인공지능법은 개인의 권리와 안전을 보호하고 인공지능 기술의 책임감 있는 발전을 도모하기 위해 범용 인공지능 모델 공급자에게 여러 의무를 부과한다. ① 기술 문서 작성 및 최신 상태 유지, ② 범용 인공지능 모델을 통합하는 인공지능 시스템 공급자에

계 정보 및 문서 제공(인공지능 모델의 기능과 한계, EU 인공지능법 상 법령 준수 사항) ③ EU 저작권법 준수 정책 마련 ④ 범용 인공지능 모델 학습에 사용된 콘텐츠에 대한 상세 요약본 작성 및 공개가 그것이다.

EU 인공지능법은 시스템적 위험성이 있는 범용 인공지능 모델 공급자에 대해서는 추가 의무를 부과하고 있는데, 여기에는 ① 표준화된 프로토콜 및 도구에 따른 모델 평가 수행(위험 식별 및 완화를 위한 적대적 테스트 수행 및 문서화 포함), ② 모델 개발·출시·사용으로 인한 시스템적 위험에 대한 평가 및 완화 조치, ③ 심각한 사고 및 대응 조치를 문서화 하여 필요시 인공지능 사무국 및 관할 당국에 보고, ④ 범용 인공지능 모델과 모델의 물리적 인프라에 대해 적절한 수준의 사이버 보안 및 보호의 보장 조치가 포함된다.

범용 인공지능 모델을 통합한 인공지능 시스템이 반드시 고위험 인공지능 시스템에 해당하는 것은 아니나, 해당하는 경우 고위험 인공지능 시스템에 대한 의무를 준수하여야 한다(section 2). 여기에는 ① 인공지능 시스템의 전체 생애 주기 동안의 인공지능 시스템 사용과 관련된 위험을 식별하고 완화할 수 있는 위험 관리 시스템 구축(테스트 절차 포함), ② 인공지능 시스템의 학습, 검증 및 테스트에 사용되는 데이터 세트가 고품질 표준을 충족해야 하며 차별적인 결과를 초래할 수 있는 편견을 포함하지 않도록 데이터 품질 관리, ③ 인공지능 시스템이 EU 인공지능법의 요건을 준수하고 있음을 입증하기 위한 포괄적인 기술 문서(technical documentation) 작성 및 기록 보관, ④ 기술적으로 인공지능 시스템 생애 주기 동안 이벤트(logs)를 자동 기록, ⑤ 인공지능 시스템의 투명성을 위해 배포자에게 시스템 작동 및 한계에 대한 정보를 제공, ⑥ 오류나 오작동을 최소화하고 수정하기 위해 적절한 인적 감독을 보장하도록 설계, ⑦ 견고하고 정확하며 적절한 수준의 사이버 보안 마련 조치가 포함된다.

그리고 EU 인공지능법은 고위험 인공지능 시스템 공급자의 의무로 ① 공급자의 신원 기재, 적합성 평가 수행, 적합성 선언서 작성 및 CE 마크 부착, EU 데이터베이스 등록(article 16), ② 품질 관리 체계 구축 및 운영(article 17), ③ 기술문서 등 문서 보관(article 18), ④ 로그 보관(article 19), ⑤ 리콜 조치 및 관할 당국에 정보 제공(article 20), ⑥ 관할 당국과의 협력(article 21) 의무를 규정하고 있고, 그 외 유통자, 배포자 등 인공지능 시스템 당사자에 대하여도 의무 사항을 규정하고 있다. 특히 자연인의 신용도를 평가하거나 신용점수를 정하기 위한 인공지능 시스템 등 일정한 경우에 고위험 인공지능 시스템 배포

자의 기본권 영향 평가 의무도 규정하고 있는데(article 27), 평가 항목에는 ① 배포자의 프로세스에 대한 설명, ② 고위험 인공지능 시스템이 사용될 기간과 빈도에 대한 설명, ③ 영향을 받을 수 있는 자연인 및 집단의 범주에 대한 설명, ④ 구체적인 피해 위험, ⑤ 인적 감독 조치의 이행에 대한 설명, ⑥ 위험이 현실화될 경우 취해야 할 조치에 대한 설명이 포함된다.

한편, EU 인공지능법은 특정 인공지능 시스템에 대해서 투명성 의무를 부과하고 있으며, 범용 인공지능 모델이 인공지능 시스템에 통합되는 경우 인공지능 시스템 공급자 등에 위 의무 사항을 부과한다. 사람과 직접 상호 작용하도록 설계된 인공지능 시스템과 관련하여서는, 인공지능 시스템 공급자가 이용자에게 인공지능 시스템과 상호작용하고 있음을 알려야 한다(article 50(1)). 오디오, 이미지, 동영상 또는 텍스트 합성 콘텐츠를 생성하는 범용 인공지능 시스템을 포함하는 인공지능 시스템의 공급자는, 인공지능 시스템의 출력값이 기계 판독이 가능한 형식으로 표시되고 인위적으로 생성 또는 조작된 것으로 감지할 수 있도록 해야 한다(article 50(2)). 인공지능 시스템 공급자는 관련 기술 표준에 반영될 수 있도록, 다양한 유형의 콘텐츠 특수성과 한계, 구현 비용 및 일반적으로 인정되는 최신 기술을 고려하여, 기술적으로 가능한 한 기술 솔루션이 효과적이고 상호 운용 가능하며 견고하고 신뢰할 수 있도록 보장해야 한다. 딥페이크에 해당하는 이미지, 오디오 또는 비디오 콘텐츠를 생성하거나 조작하는 인공지능 시스템의 배포자(공공의 관심사에 대해 대중에게 알릴 목적으로 게시되는 텍스트를 생성 또는 조작하는 인공지능 시스템의 배포자 포함)는 해당 콘텐츠가 인위적으로 생성 또는 조작되었다는 사실을 공개해야 한다(article 50(4)). 이러한 의무들은 늦어도 인공지능 시스템과의 첫 번째 상호작용 또는 인공지능 시스템의 첫 번째 노출 시점에 명확하고 구별 가능한 방식으로 이행되어야 한다.

또, 개인정보 보호 및 개인 데이터 보호에 대한 권리는 인공지능 시스템의 전체 생애 주기 동안 보장되어야 한다고 하면서(recital 69), 기본적으로 EU GDPR에 명시된 데이터 최소화 및 데이터 보호의 원칙을 인공지능 관련 개인 데이터 처리 시 적용한다. 이러한 원칙을 준수하기 위해서는, 공급자가 익명화 및 암호화 뿐 아니라 EU 인공지능법에서 규정한 데이터 거버넌스 요건을 준수하면서, 당사자 간 전송 또는 원시/구조화된 데이터 자체의 복사 없이, 데이터에 알고리즘을 적용하고 인공지능 시스템을 학습시키는 기술을 사용하는 방안도 권장된다.

나. 인공지능 고위급 전문가 그룹(AI HLEG)의 신뢰할 수 있는 인공지능을 위한 윤리 가이드라인(Ethics Guidelines for Trustworthy AI, 2019. 4. 8.)

인공지능 고위급 전문가 그룹(AI HLEG)은 신뢰할 수 있는 인공지능을 위한 윤리 가이드라인을 발표하였다. 본 가이드라인은 신뢰할 수 있는 인공지능의 세 가지 구성 요소를 합법성(모든 관련 법률 및 규정을 준수), 윤리성(윤리적 원칙과 가치를 준수), 견고성(기술적 및 사회적 관점에서 견고)이라고 보았다.

신뢰할 수 있는 인공지능을 위한 7가지 핵심 요구 사항은, ① 인간의 개입 및 감독, ② 기술적 견고성 및 안전성, ③ 프라이버시 및 데이터 거버넌스, ④ 투명성, ⑤ 다양성, 차별 금지 및 공정성, ⑥ 환경 및 사회적 복리, ⑦ 책임성이다. 본 가이드라인은 핵심 요구 사항을 실제로 구현하기 위한 평가 목록도 제시하고 있는데, 이는 주로 사용자와 직접 상호 작용하는 인공지능 시스템에 적용되며 인공지능 시스템의 개발자 및 배포자를 대상으로 한다.

다. AI HLEG의 신뢰할 수 있는 인공지능을 위한 체크리스트(The Assessment List for Trustworthy AI; ALTAI, 2020. 7. 17.)

AI HLEG는 신뢰할 수 있는 인공지능을 위한 체크리스트를 발표했다. 본 체크리스트는 인공지능 관련 개발자와 배포자가 위 인공지능을 위한 윤리 가이드라인에서 제시된 7가지 핵심 요구 사항을 실제로 구현할 수 있도록 하기 위해 개발되었다.

예를 들어 7가지 핵심 요구 사항 중 ③ 프라이버시 및 데이터 거버넌스 항목과 관련하여 살펴보면, 개인정보 보호에 대한 권리에 대한 인공지능 시스템의 영향을 고려했는지, 개인정보 보호 문제를 신고할 수 있는 메커니즘이 설정되었는지(프라이버시), 인공지능 시스템을 관련 표준(ISO 25, IEEE26 등) 또는 일일 데이터 관리 및 거버넌스를 위해 널리 채택된 프로토콜에 맞게 조정하고 있는지(데이터 거버넌스)가 고려된다. ④ 투명성과 관련하여서는, 인공지능 시스템의 결정 등을 기록하기 위한 적절한 로깅 절차가 마련되었는지(추적성), 인공지능 시스템의 결정을 사용자에게 설명했는지(설명성), 대화형 인공지능 시스템(챗봇, 로봇어드바이저 등)의 경우 사용자에게 인공지능 시스템과 상호작용하고 있다는 사실을 알렸는지(커뮤니케이션)를 검토하며, ⑤ 다양성, 차별 금지 및 공정성 항목과 관련하여서는, 데이터에서 최종 사용자 및 영향 받는 자의 다양성과 대표를 고려했는지, 편향 관련 신고 메커니즘을 구현했는지(불공정한 편향 방지), 인공지능 시스템이 사회의 다양한

선호도와 가용성에 부합하는지(접근성 및 범용 디자인), 인공지능 시스템 설계 및 개발에 가장 광범위한 범위의 이해관계자의 참여를 포함하는 메커니즘을 고려했는지(이해관계자 참여)를 검토하고 있다.

3. 미국

미국은 인공지능 거버넌스를 위한 포괄적인 연방법이 부재하지만, 인공지능 거버넌스에 관한 정부 정책 및 관행을 지시하기 위해 여러 행정 명령을 발표한 바 있다. ‘국가 AI 연구 개발 전략 계획’에서 연방정부의 인공지능 관련 연구 개발 투자를 제시했고, ‘AI 권리장전 청사진’에서 ‘안전하고 효과적인 시스템, 알고리즘 차별 보호, 데이터 프라이버시, 고지 및 설명, 인간적 대안 고려 및 대체’ 등 자동화된 시스템의 설계와 배포를 규율하는 다섯 가지 원칙을 제안하였다.

안전하고, 보안적이며, 신뢰할 수 있는 인공지능 개발과 활용을 위한 행정 명령(The Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI, Executive Order 14110)에서 위 다섯 가지 원칙에 ‘혁신과 경쟁 촉진, 근로자 지원, 연방정부의 인공지능 사용 확대, 해외에서의 미국 리더십 강화’를 추가하였으며, 수십 개의 연방 기관이 150개 이상의 조치를 취하도록 요구함으로써 대부분의 연방 기관과 연방정부와 사업을 하는 인공지능 가치 사슬 내의 기관에 바이든 행정명령이 적용되었다.

즉, 미국의 인공지능 규제 접근 방식은 연방 기관을 통한 지침 및 표준 제정과 산업 자율 규제 방식으로 요약된다. 이러한 지침에는 국립표준기술연구소(NIST)의 ‘AI RMF’, 정보통신국(NTIA)의 ‘인공지능 책임가능성 정책(AI Accountability Policy)’, 국무부(DOS)의 ‘기업 인공지능 전략(Enterprise AI Strategy)’, 국토안보부(DOS)의 ‘인공지능 로드맵(AI Roadmap)’, 관리예산처(OMB)의 ‘정부 기관의 인공지능 활용을 위한 거버넌스, 혁신, 위험 관리 촉진(Advancing Governance, Innovation, and Risk Management for Agency Use of AI)’, 재무부(DOT)의 ‘금융 서비스 분야에서 인공지능 관련 위기에 대한 보고서(Report on Managing AI-Specific Risks in the Financial Services Sector)’, 법무부(DOJ)의 ‘기업 준수 프로그램 평가(Evaluation of Corporate Compliance Programs)’ 지침, 저작권청(Copyright Office)의 ‘저작권과 인공지능(Copyright and Artificial Intelligence)’ 등이 있다.

연방 의회에서는 인공지능 훈련 법안(AI Training Act), 2020년 국가 인공지능 이니셔티

브 법안(National AI Initiative Act of 2020), 2020년 정부 내 인공지능 법안(AI in Government Act of 2020), 미국 첨단 인공지능 법안(Advancing American AI Act), CHIPS and Science Act 등 인공지능 관련 법안을 제정한 바 있다. 위 법안들은 주로 연방 정부 내에서 인공지능 채택을 촉진하고 인공지능 사용과 관련하여 연방 기관 간의 조정을 달성하는 것을 목표로 하고 있다. 2023년 알고리즘 책임가능성 법안(Algorithmic Accountability Act of 2023)이 2019년, 2022년, 2023년에 발의되어 회기 만료로 폐기된 바 있으나 2024년에도 발의되어 의회 심의를 앞두고 있다. 위 알고리즘 책임법안이 통과되면, 연방거래위원회(이하 FTC)가 소비자의 개인 정보를 포함하는 인공지능 시스템을 사용하는 기업을 규제할 수 있는 권한을 부여 받아 기업이 차별, 편견 및 보안 문제를 식별하고 해결하기 위한 영향 평가를 수행하도록 요구할 수 있게 된다. 현재 FTC를 포함한 연방 기관들은 기존 법적 권한이 인공지능을 포함한 신기술 사용에 적용된다는 점을 분명히 하고 있다.

참고로, 미국 ‘양당 상원 인공지능 작업단’(Bipartisan Senate AI Working Group)은 ‘인공지능 분야에서의 미국의 혁신 추진: 미국 상원 내 인공지능 정책을 위한 로드맵’을 마련하여 초당적 차원에서 정책적으로 우선 고려되어야 하는 8개 사항을 제시하였다. 이하에서 살펴보듯이, 여러 주에서도 인공지능 법률을 제안했으며 콜로라도(SB 205), 유타(SB 149), 캘리포니아(AB 2013, SB 942) 등 일부 주에서는 인공지능 법률을 제정하기도 했다.

다만, 트럼프 대통령 당선인은 2024년 선거 공약집에서 “인공지능 혁신을 방해하고 이 기술의 발전에 급진 좌파적 사상을 강요하는 조 바이든의 위험한 행정명령을 폐지할 것”이라고 언급한 바 있고 규제 완화와 민간 부문의 혁신에 중점을 둘 것이라고 공개적으로 발언하였으므로, 바이든 대통령이 도입한 인공지능 분야의 연방 조치를 완화할 가능성이 있다.

가. 바이든 행정명령

바이든 대통령의 행정명령 14110호는 연방 기관이 지침을 통해 국가 안보 및 소비자 보호와 같은 영역에서 위험을 평가하고 해결하도록 함으로써 인공지능에 대한 연방 차원의 조율된 접근 방식을 확립하는 것을 목표로 한다. 본 행정명령은 연방 기관을 대상으로 하지만, 명시된 원칙과 정의된 구체적인 업무는 민간 부문에 영향을 미치는 지침과 규제(특히 표준, 모델 테스트 및 차별 금지와 같은 규제)로 이어질 수 있다. 특히 본 행정명령이

구체적으로 NIST에서 개발한 인공지능 위험 관리 프레임워크(AI RMF 1.0)를 언급하고 있는 점으로 미루어 보아, 위 프레임워크가 인공지능 거버넌스의 사실상 표준이 될 수 있을 것을 시사한다.

‘인공지능’은 ‘인간이 정의한 일련의 주어진 목표에 대해 실제 또는 가상 환경에 영향을 미치는 예측, 권고 또는 결정을 내릴 수 있는 기계 기반 시스템’이라고 정의되고 ‘인공지능 시스템은 기계와 인간의 입력을 통해 현실 및 가상 환경을 인식하고 자동화된 분석을 통해 이러한 인식을 모델로 추상화하며 모델 추론을 사용하여 정보나 조치에 대한 선택지를 정립한다.’ 이는 미국연방법전 제15편 제9401조(3)에 정의를 차용하고 있는 것이다. ‘생성형 인공지능’은 ‘파생된 합성 콘텐츠를 생성하기 위하여 입력 데이터의 구조와 특성을 모방하는 인공지능 모델의 유형’이라고 정의되고 있다.

본 행정명령은 13개의 섹션으로 나뉘며, 그 중 8개 섹션에 인공지능과 관련한 원칙이 포함되어 있다. 8가지 원칙은 ① 인공지능 기술의 안전 및 보안 보장, ② 혁신과 경쟁 촉진, ③ 근로자 지원, ④ 형평성 및 시민권 증진, ⑤ 소비자, 환자, 승객 및 학생의 이익 보호, ⑥ 개인정보 보호, ⑦ 연방 정부의 인공지능 사용 증진, ⑧ 해외에서 미국의 리더십 강화이며, 연방기관 등은 위 8가지 원칙에 따라 인공지능 개발·이용을 발전시키고 관리하게 된다.

예를 들어, ① 인공지능 기술의 안전 및 보안 보장 항목과 관련하여, NIST는 인공지능 안전 및 보안을 위한 지침, 표준 및 모범사례를 개발하여야 하며, 이중 용도 기초 모델을 개발 또는 개발하고자 하는 회사는 ‘모델 훈련 정보, 모델 가중치의 소유권 및 보호 조치, 레드팀 테스트 결과 및 보안 조치 등에 관한 정보·보고서·기록’을 연방정부에 제공해야 한다.

합성 콘텐츠와 관련하여서는 합성 콘텐츠로 인한 위험 완화를 위해, 인공지능 시스템을 활용하여 생성된 합성 콘텐츠를 식별하고 라벨링하는 기능을 육성하고, 생성된 합성 및 비합성 디지털 콘텐츠의 진위성과 출처를 확증하기 위한 조치를 요구한다. 이에 NIST는 ‘생성형 인공지능 위험 관리 프레임워크: 생성형 인공지능 개요서(AI RMF Generative AI Profile)’, ‘생성형 인공지능과 이중-사용 파운데이션 모델을 위한 소프트웨어 개발 보안 지침(Secure Software Development Practices for Generative AI and Dual-Use Foundation Models)’, ‘합성 콘텐츠로 인한 위기 방지책(Reducing Risks Posed by Synthetic Content)’,

‘인공지능 표준을 위한 국제 협력 계획(A Plan for Global Engagement on AI Standards)’ 등의 가이드라인과 보고서를 발간하였으며, 생성형 인공지능 평가프로그램인 ‘NIST GenAI’도 운영 중이다.

1) NIST, 인공지능 위험 관리 프레임워크(Artificial Intelligence Risk Management Framework; 이하 AI RMF 1.0, 2023. 1. 26.)

NIST는 2020년 국가 인공지능 이니셔티브 법안(National Artificial Intelligence Initiative Act of 2020)에 따른 미 의회의 지시로 인공지능 프레임워크 개발을 준비하여 2023년 1월 26일 인공지능 위험 관리 프레임워크(AI RMF)를 발표하였다. AI RMF는 인공지능 시스템을 설계, 개발, 배포 또는 사용하는 조직을 대상으로 하며, 해당 조직이 AI RMF를 자발적으로 적용하여 인공지능 기술의 위험을 관리할 수 있도록 지원하고 있다.

AI RMF는 두 부분으로 나뉘어져 있다. 첫 번째 부분에서는 조직이 인공지능과 관련된 위험을 구성하는 방법을 논의하고 신뢰할 수 있는 인공지능 시스템의 7가지 특성을 설명한다. 두 번째 부분에서는 조직이 실제로 인공지능 시스템의 위험을 해결하는 데 도움이 되는 네 가지 구체적인 기능(관리, 매핑, 측정, 관리)을 설명한다.

신뢰할 수 있는 인공지능 시스템은 ① 타당하고 신뢰할 수 있으며, ② 안전하고, ③ 보안이 철저하고, 탄력적이며, ④ 책임을 질 수 있고, 투명하고, ⑤ 설명 및 해석이 가능하고, ⑥ 개인정보보호가 강화되고, ⑦ 유해한 편향 관리를 통한 공정한 특성을 가진다. AI RMF는 이러한 특징들이 서로 영향을 미치기 때문에 상황에 따라 맥락적으로 접근하는 것이 필요하며 특성 간 균형을 이루는 것이 중요하다고 강조한다.

이 중 ④ 책임성 및 투명성과 관련하여, ‘책임성’은 투명성을 전제이며, ‘투명성’이란 ‘인공지능 시스템과 상호작용하는 개인이(인지하고 있는지 여부와 관계없이) 해당 시스템과 그 결과물에 대한 정보를 어느 정도 접근할 수 있는지’와 관련된 특성이다. 따라서 의미 있는 투명성은 인공지능 생애 주기의 단계에 따라 적절한 수준의 정보에 대한 접근을 제공하고 인공지능 시스템과 상호 작용하거나 사용하는 인공지능 주체 또는 개인의 역할이나 지식에 맞게 조정되어야 한다.

⑤ 설명 및 해석가능성과 관련하여, ‘설명 가능성’은 ‘인공지능 시스템 작동의 근간이 되는 메커니즘을 표현하는 것’을 말하며, ‘해석 가능성’은 ‘설계된 기능적 목적의 맥락에서 인

공지능 시스템의 산출물의 의미를 나타내는 것'이다. 설명 가능성 부족으로 인한 위험은 사용자의 역할, 지식, 기술 수준 등 개인별 차이에 맞춘 설명을 통해 인공지능 시스템의 작동 방식을 설명함으로써 관리할 수 있고, 해석 가능성에 대한 위험은 인공지능 시스템이 특정 예측이나 추천을 한 이유에 대한 설명을 전달함으로써 해결될 수 있다.

⑦ 공정성과 관련하여서는, 인공지능으로 인한 편향을 인공지능 데이터세트, 인공지능 생애 주기 전반의 조직 규범, 관행 및 프로세스, 그리고 인공지능 시스템을 사용하는 광범위한 사회에서 발생하는 '시스템적 편향', 인공지능 데이터 세트와 알고리즘 프로세스에 존재하는 '계산 및 통계적 편향', 개인이나 그룹이 인공지능 시스템 정보를 인식하는 방식 또는 인공지능 시스템의 목적과 기능에 대해 인간이 생각하는 방식과 관련된 '인간 인지적 편향'으로 구분한다.

AI RMF는 신뢰할 수 있는 인공지능 시스템을 개발하기 위한 4가지 핵심 기능으로 '거버넌스-매핑(위험식별)-측정-관리'를 들고 있는데, 각 기능은 상호 연결되어 있으며 인공지능 시스템의 생애 주기 전반에 걸쳐 반복적으로 수행되어야 한다. '거버넌스' 기능은 인공지능 생애 주기 동안 인공지능 시스템이 야기할 수 있는 위험을 예측, 식별 및 관리하기 위한 프로세스와 문서 및 조직 체계를 구축함으로써 조직 내 위험 관리 문화를 조성하고 구현하는 것을 의미한다. '매핑' 기능은 인공지능 시스템과 관련된 위험을 파악하기 위한 맥락을 설정하는 것으로서, 인공지능 시스템의 위험 및 광범위한 위험 유발 요인을 식별하는 작업이다. '측정' 기능은 정량적, 정성적 또는 혼합 방법 도구, 기술 및 방법론을 사용하여 인공지능 위험 및 관련 영향을 분석, 평가, 벤치마킹 및 모니터링 하는 것이다. '관리' 기능은 거버넌스 기능에서 정의한 대로 정기적으로 매핑되고 측정된 위험을 위험 리소스에 할당하는 작업이다. 이러한 위험 처리 방법에는 사고 또는 이벤트에 대한 대응, 복구, 커뮤니케이션 계획이 포함된다.

2) NIST, 인공지능 위험 관리 프레임워크: 생성형 인공지능 개요서(Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, 2024. 7. 26.)

본 프레임워크는 바이든 행정명령을 이행하기 위해 개발되었으며, 조직이 생성 인공지능의 개발 및 사용으로 인해 발생하는 고유하거나 악화되는 12가지 고유한 위험을 식별하고 이를 완화하기 위해 취할 수 있는 조치를 AI RMF에 매핑하여 제시한다.

12가지 위협이란, ① CBRN(화학 물질, 생물학적 물질, 방사성 물질, 핵무기) 정보 또는 능력, ② 조작(confabulation), ③ 위협하고 폭력적이거나 혐오스러운 콘텐츠, ④ 데이터 프라이버시, ⑤ 환경적 영향, ⑥ 유해한 편견 또는 동질화, ⑦ 인간-인공지능 상호작용(configuration), ⑧ 정보 무결성, ⑨ 정보 보안, ⑩ 지적 재산, ⑪ 음란하고 저속하거나 모욕적인 콘텐츠, ⑫ 가치 사슬 및 구성 요소 통합이다. 본 프레임워크는 생성형 인공지능 위협을 완화하기 위해 조직이 취할 수 있는 구체적인 조치도 제시하는데, 조치들은 AIRMF의 4가지 핵심기능(GOVERN, MAP, MEASURE, MANAGE)과 관련된 하위 범주 별로 분류되며, 관련 생성형 인공지능 위협 및 인공지능 행위자의 업무와 연결된다.

예를 들어 ② 조작과 관련된 대응 조치에 대하여, 생성형 인공지능 위협 측정 개선 프로세스를 수립하고, 설명 가능성 및 투명성 부족과 관련된 일반적인 위협을 해결하기 위해 다양한 기법을 사용해야 하며(GOVERN), NIST Dioptra와 같은 특수 제작된 테스트 환경을 활용하여 생성형 인공지능의 신뢰할 수 있는 특성을 경험적으로 평가하고(MEASURE), 능동적 학습 기법을 구현하여 모델이 실패하거나 예기치 않은 출력을 생성하는 사례를 식별한다(MANAGE).

③ 위협하고 폭력적이거나 혐오스러운 콘텐츠와 관련하여서는, 법률을 위반하는 콘텐츠를 생성하지 못하도록 방지하는 정책 및 메커니즘을 수립하고(GOVERN), 조직의 위협 허용 범위를 초과하는 생성형 인공지능 시스템의 예측 가능한 불법 사용 또는 응용 프로그램을 식별하고 문서화하며(MAP), 생성형 인공지능 훈련 및 유지 관리 중에 성적으로 노골적이거나 공격적이거나 폭력적인 정보에 노출된 가치 사슬 또는 기타 인공지능 행위자에 대한 건강 및 웰빙 영향을 포함하여 악영향을 평가하고(MEASURE), 사용자가 보고한 문제 콘텐츠를 평가하고 피드백을 시스템 업데이트에 통합한다(MANAGE).

⑥ 유해한 편견 또는 동질화 항목과 관련하여서는, 특정 사고 유형에 따라 다양한 구성과 책임을 가진 팀을 생성형 인공지능 시스템 사고 대응에 참여시키는 절차를 수립하고(GOVERN), 위협 측정 및 관리 기능에 대한 정보를 제공하고 수행하기 위해 기업 전체에서 광범위한 역량, 역량, 인구 통계 그룹, 도메인 전문 지식, 교육 배경, 살아온 경험, 직업 및 기술을 반영하는 학제 간 팀을 구성하고 권한을 부여하며(MAP), 시스템적 편견을 측정하기 위해 공정성 평가를 수행하며(MEASURE), 설명할 수 없는 생성형 인공지능 시스템과 관련된 위협을 완화하기 위해 그 일환으로 설명 가능한 인공지능(XAI) 기술을 적용한다

(MANAGE).

⑦ 인간-인공지능 상호작용(configuration) 항목과 관련하여서는, 챗봇 및 의사 결정 작업을 위한 생성형 인공지능 인터페이스·방식 및 인간-인공지능 구성에 대한 허용 가능한 사용 정책을 정의하고(GOVERN), 인간 숙련도 테스트와 생성형 인공지능 기능 테스트를 구분하며(MAP), 생성형 인공지능 시스템 인터페이스에서 의인화 사례를 추적하고 문서화하며(MEASURE), 생성형 인공지능 콘텐츠 성능과 영향력에 대한 사용자 감정을 측정하기 위해 감정 분석의 사용을 중요시하고, 사용자 연구 및 경험에 숙련된 인공지능 행위자와 협력한다(MANAGE).

⑧ 정보 무결성 항목과 관련하여서는, 사용된 콘텐츠 출처 증명 방법론(암호화, 워터마킹, 스테가노그래피 등)의 효과성을 측정하기 위한 정책을 수립하고(GOVERN), 생성형 인공지능 시스템이 외부 네트워크와 상호 작용하는 방식을 관찰하고 분석하며 잠재적인 부정적인 외부 효과를 식별하고(MAP), 콘텐츠 출처를 분석하고, 데이터 이상을 감지하며, 디지털 서명의 신뢰성을 확인하고, 잘못된 정보 또는 조작과 관련된 패턴을 식별하도록 설계된 도구를 통합하며(MEASURE), 인공지능 생성 데이터의 경로와 신뢰성을 추적하고 검증하는 데 도움이 되는 실시간 감사 도구를 사용한다(MANAGE).

본 프레임워크 부록 A에는 NIST가 GAI PWG(Generative AI Public Working Group)과 협의하여 도출한 생성형 인공지능과 관련된 위험을 관리하는 데 도움이 되는 네 가지 주요 영역(거버넌스, 사전 배포 테스트, 콘텐츠 출처, 사고 공개)에 대해서 부연 설명하고 있다.

3) NIST, 합성 콘텐츠로 인한 위기 방지책(Reducing Risks Posed by Synthetic Content, 2024. 11. 20.)

NIST 산하 미국 인공지능 안전 연구소는 합성 콘텐츠로 인한 위험을 줄이기 위한 가이드라인을 발표했다. 본 가이드라인에는 콘텐츠 인증, 워터마킹, 합성 콘텐츠 탐지, 유해한 인공지능 생성 자료 방지 등 디지털 콘텐츠 투명성을 위한 자발적인 기술적 접근 방식이 자세히 설명되어 있다.

생성 인공지능 기술은 합성 콘텐츠를 생성할 수 있어 유용하지만 디지털 정보 및 커뮤니케이션의 신뢰, 안전, 투명성, 신뢰성에 대한 새로운 위험을 초래할 수 있다. 특히 합성 콘텐츠는 허위 정보, 신원 사기 또는 CSAM(child, sexual, abuse, material)과 같은 피해를

확대할 수 있다.

본 가이드라인은 디지털 콘텐츠 투명성과 관련한 기술적 접근 방식을 데이터 출처 추적(provenance data tracking) 방식과 합성 콘텐츠 탐지(synthetic content detection) 방식으로 설명한다. 데이터 출처 추적 방식에는 디지털 워터마킹과 메타데이터 기록이 해당된다. 디지털 워터마킹은 일반적으로 콘텐츠(이미지, 텍스트, 오디오 또는 동영상)에 정보를 삽입하여 제거하기 어렵게 만드는 것이다. 워터마크의 특징은 정보를 전달하기 위해 별도의 채널(예: 메타데이터 필드)을 사용하지 않고 픽셀, 단어 등을 통해 콘텐츠 자체에 정보를 인코딩하는 것이다. 워터마크는 명시적 및 비명시적 워터마킹, 프라이빗 및 퍼블릭 워터마킹, 가역적 및 비가역적 워터마킹으로 구분된다. 워터마킹이 해결해야 할 과제로는 강건성, 탐지 정확도, 용량 및 품질 문제, 보안을 들고 있다.

메타 데이터 기록 방식은 설명 정보가 콘텐츠와 연관되는 출처 데이터 추적 기법으로서, 콘텐츠의 출처, 시간 및 날짜를 나타낼 수 있다. 메타 데이터 유형에는 콘텐츠 파일 내부에 정보를 저장하는 내장 메타 데이터 유형(이미지 파일의 EXIF 정보 등)과 외부에 별도로 저장되고 해시 함수화되어 콘텐츠와 연결되는 외부 메타 데이터로 구분된다. 메타 데이터 기록 방식과 관련하여서는, 조작 및 제거, 프라이버시 위험, 표준화 문제가 해결되어야 한다.

합성 콘텐츠 탐지란 ‘주어진 콘텐츠가 인공지능 또는 기타 도구나 알고리즘에 의해 생성되었는지 여부를 분류하는 데 사용되는 기술, 방법 및 도구’를 말한다. 합성 콘텐츠 탐지 기술은 디지털 워터마크(공개 또는 은닉 방식) 또는 메타데이터 등을 통해 추적된 출처 데이터를 탐지하는 ‘출처 데이터 탐지 방식’, 생성 또는 처리 중에 남겨진 흔적(이미지 픽셀의 규칙성 또는 불일치 등)을 기반으로 생성 또는 수정된 후의 합성 콘텐츠를 식별하는 ‘자동화된 콘텐츠 기반 탐지 기술’, 전문가가 자동화된 도구를 보강하거나 보완하여 합성 콘텐츠를 식별하는 데 도움을 주는 휴먼 인 더 루프(human-in-the-loop) 방법인 ‘인간 보조 탐지’로 구분된다. 다만 이미지 및 비디오 파일의 경우 단순 편집(예: 크기 조정, 압축)만으로도 강건성이 약화 될 수 있다는 문제, 텍스트 파일의 경우 텍스트를 변용(paraphrasing)하거나 재구성하면 탐지 마커가 쉽게 가려지는 문제, 오디오 파일의 경우 노이즈 추가 또는 형식 변환으로 합성 특성이 가려질 수 있다는 문제가 제기되고 있다.

본 가이드라인은 생성형 인공지능이 CSAM 및 동의되지 않은 친밀한 이미지(non-consensual intimate imagery, 이하 NCII)를 생성하지 않도록 하는 조치로서, 훈련 데

이터 필터링(데이터셋 큐레이션 고려), 입력 데이터 필터링(키워드 탐지, 프롬프트 검열, 사용 정책 등), 출력 데이터 필터링(유해한 콘텐츠를 인식하기 위해 학습된 머신러닝 모델 활용), 해싱 처리(확인된 CSAM 및 NCII 콘텐츠를 암호화 또는 해시 함수를 사용하여 공유 데이터 베이스 저장), 데이터 출처 추적 기술(메타데이터, 워터마크 등 출처 추적 방법 활용), 레드팀 구성 및 테스트 방안 등을 제안한다.

4) NIST, 이중-사용 파운데이션 모델 오용 위험 관리(Managing Misuse Risk for Dual-Use Foundation Models, Initial Public Draft, 2024. 7. 26.)

NIST는 이중 사용 기초 모델의 오용 위험을 관리하기 위한 지침 초안을 발표하여 공개 의견 수렴 중이다. 이 지침에는 모델 오용 위험을 완화하기 위한 7가지 목표로서, ① 잠재적 오용 위험 예측, ② 오용 위험 관리를 위한 계획 수립, ③ 모델 도용 위험 관리, ④ 오용 위험 측정, ⑤ 기초 모델을 배포하기 전에 오용이 관리되는지 여부, ⑥ 배포 후 오용에 대한 정보 수집 및 대응, ⑦ 오용 위험에 대한 적절한 투명성 제공을 들고 있다.

본 가이드라인 초안에서는 오용 방지를 위한 안전장치의 예시를 제공하고 있는데, 여기에는 ① 모델 훈련 개선(훈련 데이터 필터링, 유해 출력 감소 훈련, 안전 조치 제거 방지 훈련 포함), ② 오용 시도 탐지 및 차단(알고리즘 분류기를 이용한 오용 탐지, 오용 탐지 시 조치), ③ 모델 기능 접근 제한(접근 권한 제한, 맥락 및 사용자 기반 제한, 오용 감지 시 접근 제한, 기능별 접근 제한, 내부 오용 방지), ④ 모델 가중치에 대한 접근 수준 적절성 확보(모델 가중치 공개 제한 고려, API를 통한 미세 조정 고려, 조직 내부 접근 제한), ⑤ 심각한 오용 위험 시 개발 중단이 포함된다.

5) NTIA, 인공지능 설명가능성 정책 보고서(Artificial Intelligence Accountability Policy Report, 2024. 3. 27.)

미국 상무부 산하 통신정보국(NTIA)은 바이든 행정명령을 이행하기 위해 2024년 3월 27일 인공지능 책임 정책 보고서를 발표했다. 이 보고서의 핵심은 관련 주체들이 인공지능 시스템의 개발과 배포에 책임감을 갖도록 하기 위해 '지침', '지원', '규제 요건'의 세 가지 영역에 걸쳐 인공지능 책임에 대한 권장 사항을 제시하는 것이다.

본 보고서는 연방 정부 기관이 이해관계자들과 협력하여 기존 또는 새로운 권한을 활용하여 인공지능 감사 및 감사인을 위한 지침을 마련해야 한다고 하고, NIST의 인공지능 관

런 지침 개발이 여기에 포함된다고 명시하고 있다. 연방 정부 기관이 이해관계자들과 협력하여 표준 정보 공개 및 접근성을 개선하여 인공지능 모델, 학습 데이터, 입력 및 출력 데이터, 테스트 등에 관한 투명성을 촉진해야 한다고 자세히 설명하면서, 연방 정부 기관이 이해관계자들과 협력하여 책임 규정 및 표준을 인공지능 시스템에 적용하고 책임 소재와 책임자를 결정하기 위한 권고안을 만드는 방안도 제안한다.

본 보고서는 연방 정부 기관이, 미국 인공지능 안전 연구소에 대한 추가 지원, 국립 인공지능 연구 자원의 설립 및 자금 지원, 기타 정책 이니셔티브 등을 통하여 인공지능 시스템에 대한 독립적인 평가 체계가 마련되도록 해야 하고, 인공지능 시스템이 언제 사용되는지, 어떤 자료로 학습되었는지, 어떤 능력과 한계를 가지고 있는지 평가할 수 있는 신뢰할 수 있고 널리 적용 가능한 도구도 개발해야 한다고 권고한다.

마지막으로, 본 보고서는 연방 정부 기관이, 기존 및/또는 새로운 권한을 사용하여 고위험 인공지능 모델 부류 및 시스템에 대한 독립적이고 규제적인 검사를 요구하고, 경우에 따라 평가를 의무화하며, 인공지능 사고·배포·감사에 대한 레지스트리를 유지함으로써 인공지능과 관련된 부문 간 위험 및 관행을 해결할 수 있는 역량을 강화하고, 정부 공급업체, 계약자 및 보조금 수혜자와 관련하여 사용되는 인공지능에 대해 건전한 인공지능 거버넌스 및 보증 관행을 갖추는 조치를 취해야 한다고 한다.

나. 미국 주별 인공지능 법령

1) 유타 주

유타주는 주정부 차원으로는 최초로 2024년 3월 13일 ‘생성형 인공지능 이용 고지 및 인공지능 기술 개발을 위한 규제완화’를 위해 ‘인공지능 수정법」(Artificial Intelligence Amendments, SB0149)’을 제정하였다.

이 법에 의하면, 생성형 인공지능을 이용하여 서비스를 제공하는 경우 이를 소비자에게 고지하여야 한다. ‘규제 직종’(regulated occupation) 종사자는 생성형 인공지능을 이용하여 서비스를 제공하는 경우, ① 구두로 대화 시 대화를 시작할 때 구두로, ② 서면 소통 시 소통 전 전자메시지로 ‘생성형 인공지능을 이용함’을 상대방에게 고지(사전 공개)하여야 한다. 소비자보호 분야에서 사람과 상호작용하는데 생성형 인공지능을 사용하는 자는 소비자의 요청이 있는 경우 ‘생성형 인공지능과 상호작용하고 있다는 점’을 명확하고 분명하게 고지하여야 한다.

2) 테네시 주

테네시주는 2024년 3월 21일 ‘초상·음성·이미지 보호 보장법(Ensuring Likeness Voice and Image Security Act of 2024, SB2096)(소위 ELVIS법)’을 제정하였다. 이 법은 테네시주 개인 권리 보호법을 업데이트하여 작곡가, 연주자, 음악 산업 종사자의 목소리를 인공지능의 오용으로부터 보호하는 내용을 포함하는 것을 목표로 한다.

이 법에서 ‘음성’은 ‘실제 개인의 음성 또는 이와 유사한 음성을 포함하는지에 상관없이, 특정 개인에 속하거나 특정 개인을 쉽게 식별할 수 있는 매체 음성’이라고 정의되어 생성형 인공지능 등 기술 발전 및 이용 현실을 반영하였다. 그리고 개인의 재산권(property right)에 ‘음성’을 명시적으로 포함하여 재산권 보호 대상을 확대하였다. 이에 사전 동의 없이 음성 등을 상업적으로 이용하거나, 권한 없이 배포·전송 등을 한 자는 민사상 책임을 부담하게 되는데, 여기에는 사진·음성·초상을 배포·전송·공개하는 것을 개인으로부터 승인받지 않은 것을 알면서, 해당 특정 개인을 식별 가능한 사진·음성·초상을 만들어 내는 것을 주된 목적·기능으로 하는 알고리즘소프트웨어·도구기기를 배포·전송·공개한 자의 민사상 책임도 포함된다.

3) 콜로라도 주

콜로라도 주는 2024년 5월 17일 ‘인공지능 시스템과 상호작용을 하는 소비자 보호에 관한 법’(concerning consumer protections in interactions with artificial intelligence systems, SB205) 을 제정하였다.

위 법에서는 ‘알고리즘 차별’을 ‘인공지능 사용으로 인해, 나이, 피부색, 장애, 민족, 유전 정보, 영어 유창성 수준, 국적, 인종, 종교, 생식건강(生殖健康), 성별, 재향군인 또는 그 밖의 주·연방 법에 따라 보호되는 부류에 속하는 개인·집단에 불리한 영향 또는 불법적 차별 대우가 야기되는 모든 조건’이라고 정의하며, ‘고위험 인공지능 시스템’을 ‘배포된 후 중대한 결정을 하거나 이러한 결정에 있어서 상당한 요인을 만드는 인공지능 시스템’이라고 정의한다.

위 법은 고위험 인공지능 시스템 개발자의 의무로, ① 알고리즘 차별 위험으로부터 소비자 보호 조치, ② 위험·데이터 등에 관한 정보 제공, ③ 영향평가 수행에 필요한 정보 등 제공, ④ 시스템 유형 및 알고리즘 차별성 관리 방법 등 제공, ⑤ 차별 발생 보고서 제

공 의무를 규정한다. 고위험 인공지능을 이용하는 배포자의 의무로는 ① 알고리즘 차별 위험으로부터 소비자 보호 조치, ② 위험관리체계 및 프로그램 수립(NIST RMF, ISO/IEC 42001 표준 참고), ③ 영향평가 수행, ④ 피해 배포 시스템 검토, ⑤ 소비자에 대한 고지 및 정보 제공, ⑥ 웹사이트에 정보 게시, ⑦ 배포 후 차별 보고 의무를 규정한다. 인공지능 시스템 개발자와 배포자는 소비자가 상호작용하는 것이 인공지능 시스템이라는 것을 소비자가 알 수 있게 해야 하는 의무를 공통으로 부담한다.

4) 캘리포니아 주

캘리포니아 주는 2024년 9월 다수의 인공지능 규제 법안들을 제정하였다. ‘생성형 인공지능: 훈련 데이터 투명성(Generative artificial intelligence: training data transparency; AB 2013)’ 법안은 인공지능 개발자가 관련 웹사이트에 생성형 인공지능 시스템 또는 서비스를 학습시키는 데 사용된 데이터를 설명하는 문서를 제공하도록 요구하고 있고, ‘캘리포니아 인공지능 투명성(California AI Transparency Act; SB 942)’ 법안은 매월 100만 명 이상의 방문자 또는 사용자가 있는 인공지능 시스템 개발자에게 인공지능 콘텐츠 탐지 도구와 인공지능 워터마크 표시 기능을 제공하도록 요구하고 있으며, ‘전자통신: 자동화된 전화 안내 장치: 인공 음성(Telecommunications: automatic dialing and announcing devices: artificial voices; AB 2905)’ 법안은 인공 음성을 사용하는 모든 사전 녹음된 메시지는 최초 자연스러운 음성 안내 시 수신자에게 이 사실을 알려야 한다고 한다.

‘범죄: 성적 이미지 배포(Crimes: distribution of intimate images; SB 926)’ 법안은 인공지능이 생성한 누드 이미지로 누군가를 협박하는 것을 범죄 행위로 규정하고 있고, ‘성착취 디지털 이미지(Sexually explicit digital images; SB 981)’ 법안은 소셜 미디어 플랫폼이 딥페이크 누드에 대한 신고 채널을 구축하도록 의무화한다. ‘캘리포니아 소비자 프라이버시: 개인정보 법안(California Consumer Privacy Act of 2018: personal information; AB 1008)’에서는 개인정보가 인공지능 시스템에 저장된 정보를 포함하여 다양한 형태로 존재할 수 있음이 명시되었다.

그리고 ‘허위 딥페이크로부터 민주주의를 보호하기 위한 법안(Defending Democracy from Deepfake Deception Act of 2024; AB 2655)’에 따라 대형 온라인 플랫폼은 선거와 관련된 인공지능 딥페이크 콘텐츠를 삭제하거나 라벨을 지정하고 해당 콘텐츠를 신고할 수

있는 채널을 만들어야 하고, ‘선거법: 선전에서 기만적인 미디어 사용(Elections: deceptive media in advertisements; AB 2839)’ 법안은 선거에 대해 유권자를 속일 수 있는 인공지능 딥페이크를 게시하거나 재게시하는 소셜 미디어 사용자를 규제하고 있으며, ‘1974년 정치 개정법: 정치적 선정: 인공지능(Political Reform Act of 1974: political advertisements: artificial intelligence; AB 2355)’ 법안은 인공지능이 생성한 정치 광고에 대한 공개를 의무화하고 있다.

참고로, 프론티어 인공지능 모델에 대한 안전하고 보안적인 혁신(Safe and Secure Innovation for Frontier Artificial Intelligence Models Act; SB 1047) 법안은 포괄적인 안전 및 보안 프로토콜 구현, 인공지능 모델 출시 전 위험 평가, 정기적인 제3자에 의한 감사 등의 내용을 포함하고 있었는데, 캘리포니아 주지사가 이에 대하여 거부권을 행사하여 현재 추가 검토를 위해 캘리포니아 상원에 계류된 상황이다.

4. 영국

영국은 2021년 9월 22일 ‘국가 인공지능 전략(National AI Strategy)’을 공개하여 인공지능 강국을 목표로 한 10년의 계획을 발표하였다. 영국은 인공지능 규제에 있어서 포괄적인 법률 제정보다는 기존 법률 체계를 인공지능 분야에 적용하고, 혁신을 촉진하는 방향으로 규제를 점진적으로 도입하는 전략을 추진하고 있다. 2023년 3월경 발표된 ‘친혁신적 인공지능 규제 접근(A pro-innovation approach to AI regulation)’ 백서는 영국 인공지능 규제 정책의 핵심 방향을 제시했다.

영국은 새로운 인공지능 규제 기관을 설립하는 대신, 데이터 보호, 경쟁 및 시장, 의약품 및 의료 등 다양한 분야를 담당하는 기존 규제 기관의 역할을 강화하여 인공지능 규제를 효율적으로 수행하고자 한다. 이에 각 규제 기관은 인공지능 기술의 특성과 각 분야의 특수성을 고려하여 맞춤형 규제 지침을 개발하고 있다. 영국 정보위원회 사무소(Information Commissioner’s Office)는 2023년 3월 ‘인공지능과 데이터 보호에 대한 지침(Guidance on AI and Data Protection)’을 업데이트했으며, 2024년 1월에는 ‘생성형 인공지능과 데이터 보호에 대한 지속 협의(Consultation Series on Generative AI and Data Protection)’ 자문도 진행했다. 경쟁 및 시장 당국(Competition and Markets Authority)은 2023년 9월 인공지능

기반 모델에 대한 초기 보고서를 발표하였고, 의약품 규제청(Medicines and Healthcare Regulatory Agency)은 2023년 10월 ‘의료장치로써 인공지능과 소프트웨어(Software and AI as a Medical Device)’을 발표했으며, 가스·전력시장국(Office of Gas and Electricity Markets)과 항공청(Civil Aviation Authority)도 2024년 하반기에 인공지능 전략 발표를 준비하고 있다. 안전보건청(Health and Safety Executive), 평등인권위원회(Equality and Human Rights Commission), 커뮤니케이션청(Office of Communications), 금융행위감독청(Financial Conduct Authority)도 각 부문에서 인공지능 사용에 관한 가이드라인을 발표해 나갈 예정이다.

가. DSIT의 친혁신적 인공지능 규제 접근(A pro-innovation approach to AI regulation, 2023. 3. 29.)

영국 과학·혁신·기술부(Department for Science, Innovation and Technology)는 인공지능 위험 대응과 인공지능 혁신 도모를 위한 “친혁신적 인공지능 규제 접근이라는 ‘인공지능 규제 백서(정책 보고서)’를 마련하였다. 위 백서는 ① 책임있는 혁신 도모 및 규제 불확실성 감소를 통한 성장과 번영 추진, ② 위험 규율과 기본가치 보호를 통해 인공지능에 대한 공공 신뢰 제고, ③ 인공지능 주도 국가로서 입지 강화를 목표로 한다.

위 백서는, 인공지능 규제 프레임워크는 ① 안전, 보안 및 견고성, ② 적절한 투명성 및 설명 가능성, ③ 공정성, ④ 책임성·거버넌스, ⑤ 이의제기 및 구제 가능성 등 5가지 부문 간 구속력 없는 원칙에 기반해야 한다고 설명하고, 프레임워크 이행 지침에 대한 주관은 과학·혁신·기술부가 담당하고 ICO 등 개별 규제기관이 위 5가지 원칙에 따라 프레임워크를 구체적으로 이행하여야 한다고 한다.

이에, 규제기관들은 ① 안전, 보안 및 견고성 원칙과 관련하여서는 기술적으로 안전한 인공지능을 구현하고 인공지능이 생애 주기를 통해 의도된 기능을 수행하게 할 수 있도록 지원하는 지침을 마련해야 하고, ② 적절한 투명성 및 설명 가능성 원칙과 관련하여서는 규제지침 마련 등 투명성 조치 지원·이행 방안을 마련하며 인공지능의 영향을 받는 자의 권리 행사에 필요한 정보 접근이 보장되어야 하며, ③ 공정성 원칙과 관련하여서는 소관 분야 공정성 설명자료 마련·공개 및 관련 법령·표준 등에 관한 지침 개발을 해야 하고, ④ 책임성·거버넌스 원칙과 관련하여서는 규제 이행 및 선례 안착 방법을 강구하고 거버

넌스 절차 이행을 장려하며, ⑤ 이의제기 및 구제 가능성 원칙과 관련하여서는 이의제기
·구제절차 명확화 및 이의제기 보장을 위한 적절한 조치 이행을 보장해야 한다.

나. ICO의 인공지능과 데이터 보호에 대한 지침(Guidance on AI and Data Protection, 2023. 3. 15. updated)

ICO는 2023년 3월 15일 인공지능 및 데이터 보호에 관한 지침을 업데이트했다. 이번 개정은 인공지능의 공정성 요건을 명확히 하고, 기술 도입과 개인 보호의 균형을 맞추기 위한 ICO의 3개년 전략 계획 'ICO 25'에 부합하게 하였으며, 인공지능 규제에 대한 영국 정부의 친혁신 입장과의 합치시켰다. 본 가이드라인은 개인정보 보호 법령 준수 담당자(DPO, 법률 고문, 위험 관리자, 고위 경영진, ICO 감사관, 기술 전문가(머신러닝 개발자 및 데이터 과학자, 소프트웨어 개발자/엔지니어, 사이버 보안 및 IT 리스크 관리자)를 대상으로 작성되었다.

본 가이드라인은 인공지능 및 데이터를 ① 책임성, ② 투명성, ③ 적법성, ④ 정확성, ⑤ 공정성, ⑥ 보안, ⑦ 개인의 권리 측면에서 분석하고 각 분야에서의 대응 조치를 탐구하고 있다.

예를 들어, ② 투명성 항목과 관련하여, 인공지능 시스템에서 개인 데이터를 처리하기 전에 개인 데이터를 처리하려는 개인에 대한 투명성 의무를 고려해야 하는데, '개인 데이터 처리 목적, 해당 개인 데이터의 보유 기간, 그리고 공유 대상'을 포함해 고려해야 한다. ⑤ 공정성 항목과 관련하여서는, '공정성'이란 '사람들이 합리적으로 기대할 수 있는 방식으로만 개인 데이터를 처리하고, 그들에게 부당한 부정적 영향을 미칠 수 있는 방식으로 사용하지 않아야 한다'는 의미이다. 인공지능 상황에서 데이터 보호 공정성은, 설계 단계와 인공지능 생애 주기 전반에 걸친 데이터 보호, 데이터 보호 영향 평가, 합법성, 투명성, 목적 제한, 데이터 최소화 및 저장 제한, 보안, 책임, 정확성, 프로파일링 및 자동화된 의사결정에서의 안전장치, 개인의 권리 준수에 의해 보장된다.

다. ICO의 생성형 인공지능과 데이터 보호에 대한 지속 협의(Consultation Series on Generative AI and Data Protection, 2024. 1. 15. - 2024. 9. 18.)

ICO는 올해 초 생성형 인공지능 모델의 개발과 사용에 데이터 보호법이 어떻게 적용되어야 하는지에 대한 일련의 자문 결과를 발표했다.

첫 번째 자문은 웹에서 데이터를 스크랩하거나 웹에서 스크랩한 데이터를 처리하여 생성형 인공지능 모델을 학습시키는 경우의 법적 근거에 관한 것이었다. ICO는 인공지능 모델 학습을 목적으로 개인 데이터를 수집하고 처리할 수 있는 유일한 합법적 근거는 정당한 이익이라는 견해다. 합법적 근거로서 정당한 이익을 확립하기 위해 인공지능 개발자는 ‘세 가지 테스트(목적 테스트, 필요성 테스트, 균형성 테스트)’를 통과해야 한다. 그리고 균형성 테스트와 관련해서는 원 개발자에 의한 배포, API를 통한 타사 배포, 오픈 소스 접근 방식 각각에서 위험을 완화할 수 있는 조치를 취해야 한다.

두 번째 자문은 목적 제한 원칙에 관한 것이다. 생성형 인공지능 생애 주기의 각 단계에 대해 충분히 구체적이고 명시적이며 명확한 목적을 설정해야 하고, 각 단계에서 처리되는 개인 데이터의 종류와 명시된 목적을 달성하기 위해 필요한 이유를 설명해야 한다.

세 번째 자문은 데이터와 정확성에 관한 것이다. ICO는 UK GDPR에 명시된 정확성 원칙과 인공지능 시스템 자체의 정확성을 의미하는 ‘통계적 정확성’ 개념의 차이점을 강조한다. ICO는 개인 데이터가 항상 정확할 필요는 없다고 지적하고, 생성형 인공지능 모델의 결과물이 반드시 통계적으로 100% 정확할 필요는 없으며, 정확도는 애플리케이션의 사용 목적에 따라 달라질 수 있음을 분명히 하고 있다. 다만 개인에게 영향을 미치는 결정을 내리는 애플리케이션은 그렇지 않은 애플리케이션보다 본질적으로 더 정확해야 한다.

네 번째 자문은 생성형 인공지능의 훈련 및 미세 조정 단계에서의 개인의 권리에 관한 것이다. 즉, 학습 및 미세 조정 단계에서 UK GDPR에 명시된 개인의 권리를 인공지능 생애 주기에 반영하는 것이 중요하며, 여기에는 정보를 고지 받고, 부정확한 데이터를 수정하며, 자동화된 결정을 피하고, 데이터를 삭제하며, 데이터 사용을 제한할 수 있는 권리가 포함된다. 조직은 웹 스크랩 데이터에 대한 구체적인 조치와 함께 투명성을 제공하고 권리 행사를 가능하게 해야 한다.

다섯 번째 자문은 생성형 인공지능 공급망 전반에서 데이터 보호 규정 준수에 관한 책임 분배에 대하여 다룬다. ICO는 공급망의 다양한 주체 간에 상호 의존성이 증가하고 있다고 지적하며, 자신이 컨트롤러인지, 공동 컨트롤러인지, 프로세서인지(그리고 그에 따른 책임이 무엇인지) 신중하게 고려하고 분석 결과를 적절하게 문서화(처리 활동 기록 및 데이터 보호 영향 평가 등)해야 한다고 강조한다.

라. Ofcom의 딥페이크 대비책(Deepfake Defences, 2024. 7. 23.)

Ofcom은 딥페이크에 대한 토론 문서(discussion paper)를 발표했다. 본 토론 문서는 딥페이크가 야기할 수 있는 위해로, 모욕적 콘텐츠 및 사기·허위 콘텐츠를 통해 돈을 갈취하거나 추가적인 성적 콘텐츠를 공유하도록 강요하는 데 사용될 수 있다는 점, 다른 사람의 신원을 허위로 표시하여 사기·허위 광고·로맨스 스캠에 사용될 수 있는 점, 여론에 영향을 미치는 허위 정보를 유포하는 데 사용될 수 있는 점을 들고 있다.

이 토론 문서는 위와 같은 위험을 해결하기 위해 네 가지 조치를 설명한다. ① 프롬프트 필터를 사용하여 특정 유형의 콘텐츠가 생성되지 않도록 방지, 모델 학습 데이터셋에서 유해한 콘텐츠 제거, 유해한 콘텐츠가 생성되는 것을 자동으로 차단하는 출력 필터를 사용하는 것(예방), ② 콘텐츠에 눈에 띄지 않는 워터마크를 삽입하여 딥러닝 알고리즘으로 감지할 수 있도록 하고, 콘텐츠가 생성될 때 메타데이터를 첨부하며, 업로드 시 인공지능이 생성한 콘텐츠에 자동으로 눈에 보이는 레이블을 추가하는 것(임베딩), ③ 널리 알려진 딥페이크 콘텐츠에 대해 학습된 머신러닝 분류기를 사용하는 등(정보가 첨부되었는지 여부에 관계없이) 콘텐츠의 출처를 밝히는 도구를 사용하는 것(탐지), ④ 온라인 서비스에서 생성 및 공유할 수 있는 합성 콘텐츠 유형에 대한 명확한 규칙을 설정하고 이를 위반하는 사용자에게 대해 사용자 계정을 일시 정지하거나 삭제하는 등의 조치를 취함으로써 단속을 강화하는 것(집행)이 그것이다.

마. Ofcom, 생성형 인공지능 유해에 대한 레드 팀 운영(Red Teaming for GenAI Harms, 2024. 7. 23.)

Ofcom은 온라인 안전을 보장하기 위해 생성형 인공지능 모델의 취약점을 식별하는 방법을 설명하는 토론 문서(discussion paper)를 발표했다. 본 토론 문서에서는 유해한 콘텐츠로부터 사용자를 보호하기 위해 생성형 인공지능 모델의 취약점을 찾기 위한 평가 방법의 일종인 ‘레드 팀’에 대해 주로 논의한다.

토론 문서에서는 레드팀 구성이 네 가지 단계로 구성된다고 강조한다. ① 먼저 팀을 구성하고 명확한 목표를 설정하고, ② 일련의 공격 프롬프트를 테스트 중인 인공지능 모델에 입력하며, ③ 인공지능 모델의 결과를 분석하여 어떤 공격이 유해한 콘텐츠를 생성하는지 파악하고, ④ 마지막으로 결과에 따라 조치(모델 안전 교육, 안전 조치 업데이트, 추

가 테스트 범위 조정, 모델 배포 중단 등)를 취하는 것이다.

다만, 분석하기 어려운 다양한 콘텐츠를 생성하는 비디오, 오디오 및 멀티모달 모델에서는 레드 팀 작업이 어렵다는 한계가 있는데, 토론 문서에서는 레드팀 구성의 효과를 극대화 하기 위한 모범 사항을 권고한다. 평가 대상의 피해를 명확하게 정의하고 결과를 측정하기 위한 메트릭을 설정하는 것, 다양한 레드 팀 구성원을 구성하는 것, 반복적인 평가 수행, 필요에 맞는 리소스를 제공하는 것, 추가 안전장치를 구현할 준비가 되어 있고 모델 출시를 중단할 수 있는 옵션을 보유할 것, 레드 팀을 유일한 평가 방법으로 의존하지 말 것, 결과를 문서화하고 널리 공유할 것, 레드 팀에 대한 연구를 최신 상태로 유지할 것 등이 그 권고 사항이다.

5. 캐나다

캐나다는 2017년경 범캐나다 인공지능 전략(Pan-Canadian AI Strategy)를 시작하여 인공지능 연구, 인재 및 상업화를 지원해 왔다. 캐나다는 2019년경에는 세계 최초의 인공지능 관련 법적 구속력 있는 규정인 ‘자동화된 의사 결정에 대한 지침(Directive on Automated Decision-making)’을 발표하기도 했다. 이 지침은 ‘위험 기반 접근 방식’을 사용하여 정부의 인공지능 사용을 규제하고 정부에 알고리즘 영향 평가 도구를 사용하도록 요구한다.

이처럼 캐나다는 EU AI Act와 유사한 위험 기반 프레임워크를 기반으로 인공지능을 규제하고 있다. 캐나다 정부는 2022년에 Bill C-27 법안(Digital Charter Implementation Act, 2022)의 Part III로 인공지능과 데이터 법안(AI and Data Act, 이하 AIDA)을 상정하였다. AIDA는 인공지능 시스템, 특히 ‘고영향 인공지능 시스템(high-impact AI systems)’에 대한 규제 프레임워크를 수립하고 있다.

캐나다 인공지능 거버넌스에는 다양한 측면에서 적용되는 다수의 법령이 있다. 데이터 프라이버시 및 보호 측면에서는, 디지털 헌장 이행 법안(Digital Charter Implementation Act, 2022)에서 기존의 개인정보 보호 및 전자 문서법(PIPEDA)을 디지털 경제에 맞게 소비자 개인정보 보호법(CPPA)으로 개정하는 내용이 포함되어 있다. CPPA 법안에 의하면 조직은 개인에게 중대한 영향을 미치는 자동화된 시스템에 의해 이루어진 예측, 권장 사항 또는 결정을 설명하여야 한다. 지식 재산 측면에서는, 저작권법 개정을 통해 인공지능 및

지식 재산 문제를 해결하려는 시도를 하고 있다. 여기에는 텍스트 및 데이터 마이닝(TDM), 저작권 및 소유권, 책임을 포함하여 생성형 인공지능 시스템으로 인해 발생하는 문제가 포함되어 있다. 인권 측면에서는 연방 및 주의 인권법이 중요한 역할을 하고 있으며, 소비자 보호 측면에서는 캐나다 소비자 제품 안전법(Canada Consumer Product Safety Act)과 개별 주별 소비자 보호법이 기술 중립성을 유지하면서 허위 진술 및 부당한 강요에 대해서 다루고 있다. 공정 경쟁 측면에서는, 캐나다 경쟁청(Competition Bureau of Canada)에서 2024년 5월 인공지능이 시장 경쟁에 미칠 수 있는 영향에 대한 논의를 담은 보고서를 발간하였고, 구체적으로 캐나다 소비자 제품 안전법 등 기존 법령이 위와 같은 시장 경쟁 문제 해결에 적용될 수 있다.

가. 인공지능과 데이터 법안(Artificial Intelligence and Data Act, 이하 AIDA)

캐나다는 정부는 2022년 6월 16일 디지털 헌장(Digital Charter)을 이행하기 위해 ‘디지털 헌장 이행 법안(Digital Charter Implementation Act, Bill C-27)’을 발의하였고, 해당 법안의 Part III 일부가 AIDA 법안이다. AIDA Companion document에 의하면, AIDA에 따른 의무 이행을 위해 인적 감독 및 모니터링, 투명성, 공정성 및 형평성, 안전성, 책임, 타당성 및 견고성을 원칙으로 제시한다.

AIDA는 ① 인공지능 시스템을 설계, 개발 또는 사용하기 위한 목적으로 인간 활동과 관련된 데이터를 처리하거나 활용할 수 있도록 제공하는 행위와 ② 인공지능 시스템을 설계, 개발 또는 사용할 수 있도록 제공하거나 그 운영을 관리하는 행위를 규제 대상 행위로 정하고 있다. 규제 대상 행위를 수행하고 그 행위 과정에서 익명화된 데이터를 사용할 수 있도록 하는 사람은 데이터 익명화 방식과 익명화된 데이터의 사용 또는 관리에 관한 조치를 수립해야 한다. 그리고 익명화, 완화 조치, 규정 준수 모니터링 조치와 관련된 모든 조치에 대하여 기록해야 하고 ‘고영향 인공지능 시스템’인지 여부에 대한 자체 평가 자료를 보관해야 한다.

‘고영향 인공지능 시스템’은 ‘하위 법령에서 정하는 기준을 충족하는 인공지능 시스템’이라고 정의된다. AIDA Companion document은 인공지능 시스템이 영향력이 큰 시스템인지 여부를 판단하는 데 유용한 요소로, 의도된 목적과 의도하지 않은 잠재적 결과에 근거하여 건강과 안전에 해를 끼치거나 인권에 악영향을 미칠 위험, 잠재적 피해의 심각성, 사용

규모, 이미 발생한 피해 또는 부정적 영향의 성격, 현실적 또는 법적 이유로 해당 시스템에서 오프 아웃하는 것이 합리적으로 불가능한 정도, 경제적 또는 사회적 상황의 불균형 또는 영향을 받는 사람의 연령, 다른 법률에 따라 해당 위험이 적절히 규제되는 정도를 들고 있다.

AIDA는 고영향 인공지능 시스템 책임자로 하여금 시스템 사용으로 인해 발생할 수 있는 피해 또는 편향된 결과의 위험을 식별, 평가 및 완화하기 위한 조치를 수립하도록 요구하며, 이러한 조치의 준수 여부를 모니터링 하는 조치도 수립하여야 한다고 한다. 그리고 고영향 인공지능 시스템을 제공하는 자와 그 운영을 관리하는 자는 공개적으로 이용 가능한 웹사이트에 '시스템의 사용 목적, 생성하려는 콘텐츠의 유형과 생성하려는 의사 결정, 권장 사항 또는 예측, 수립된 완화 조치, 기타 규정에서 규정하는 기타 정보'를 게시하여야 한다.

나. 캐나다 혁신과학경제개발부(ISED)의 첨단 생성 인공지능 시스템의 책임 있는 개발 및 관리에 관한 자발적 행동강령(Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems, 2023. 9. 27.)

캐나다 혁신과학경제개발부(ISED)는 첨단 생성 인공지능 시스템의 책임 있는 개발 및 관리에 관한 자발적 행동강령을 발표했다. 본 행동강령은 범용 생성 인공지능 시스템을 개발하고 관리하는 조직이 해당 시스템으로 인한 위험을 완화하기 위한 자발적 조치를 제시한다. 본 강령은 첨단 생성 인공지능 시스템 뿐 아니라 고영향 인공지능 시스템에도 광범위하게 적용될 수 있으므로, 강령 상의 조치는 AIDA에 명시된 고영향 인공지능 시스템의 의무 이행과 관련하여 참고할 수 있다.

첨단 생성 인공지능 시스템 개발자와 관리자는 ① 개발 또는 관리하는 시스템과 관련하여 자신의 역할을 이해하고, 적절한 위험 관리 시스템을 구축하며, 필요에 따라 다른 기업과 정보를 공유하여 공백을 방지해야 하고(책임성), ② 인공지능 시스템을 위험 평가의 대상으로 설정하며, 안전한 운영을 보장하기 위해 필요한 완화 조치를 배포 전에 마련해야 하고(안전), ③ 시스템 개발 및 배포의 여러 단계에서 공정성 및 형평성과 관련된 잠재적 영향을 평가하고 해결해야 하며(공정성 및 형평성), ④ 소비자가 정보에 입각한 결정을 내리고 전문가가 위험이 적절히 해결되었는지 평가할 수 있도록 충분한 정보가 공개되어야

하고(투명성), ⑤ 배포 후 시스템 사용을 모니터링하고 구체화되는 위험을 해결하기 위해 필요에 따라 업데이트를 구현해야 하며(인적 감독 및 모니터링), ⑥ 시스템이 의도한 대로 작동하고 사이버 공격으로부터 안전하며 노출될 수 있는 다양한 업무 또는 상황에 대한 시스템 동작을 이해하여야 한다(타당성 및 견고성).

다. 캐나다 사이버 보안 센터(CCCS)의 생성형 인공지능(Generative artificial intelligence; ITSAP.00.041, 2023. 7. 14.)

캐나다 사이버 보안 센터(CCCS)는 생성형 인공지능 사용에 대한 가이드라인을 발표했다. 본 가이드라인에서는 생성형 인공지능의 위험으로, 오류 정보 및 허위 정보, 피싱, 데이터의 개인정보 보호, 악성 코드, 버그 코드, 포이즌 데이터셋, 편향된 콘텐츠, 지적 재산 피해를 들고 있다.

본 가이드라인은 위험 완화 조치로, 조직 차원에서는 ① 강력한 인증 메커니즘 구현, ② 보안 패치 및 업데이트 적용, ③ 위협 및 취약성에 대한 최신 정보 유지, ④ 네트워크 보호, ⑤ 직원 교육을 취할 수 있고, 개인 차원에서는 ① 콘텐츠 확인, ② 기본 사이버 보안 실천, ③ 온라인에 노출하는 개인정보의 양을 줄이고 출처를 알 수 없는 링크를 클릭하지 않는 등 기본적인 온라인 안전 수칙을 실행하는 것을 든다.

본 가이드라인은 프라이버시 우려를 완화하면서 양질의 신뢰할 수 있는 콘텐츠를 생성하는 데 도움이 되는 방안으로, ① 생성형 인공지능 사용 정책 수립, ② 학습 데이터셋의 신중한 선택, ③ 보안이 잘 구현된 공급업체 선택, ④ 민감한 데이터 제공 금지 및 사용자에 의한 프롬프트 기록 삭제 허용을 든다.

6. 일본

일본은 인공지능 개발이나 사용을 직접적으로 금지하거나 제한하는 법률이나 규제는 없지만, 정부는 기업을 위한 윤리적 원칙과 자발적인 가이드라인을 발표하고 있다. EU 등과 달리 일본은 민간 부문의 자발적인 노력을 강조하는 방식을 취하고 있다. 인공지능 개발 또는 사용과 관련하여 지식 재산권법, 개인정보 보호법, 민법 및 형법과 같은 기존 법률이 적용될 수 있다.

일본 정부는 2019년 3월경 '인간 중심 인공지능의 사회적 원칙(Social Principles of

Human-Centric AI'을 발표했다. 위 원칙은 인공지능의 적절한 사용을 통해 추구해야 할 기본 철학으로 '존엄성', '다양성', '포용성 및 지속가능성'이라는 세 가지 가치를 제시하고, 7가지 사회 원칙으로, ① 인간 중심 원칙, ② 교육 및 문해력 원칙, ③ 개인정보 보호 원칙, ④ 보안 보장 원칙, ⑤ 공정 경쟁 원칙, ⑥ 공정성, 책임성 및 투명성 원칙, ⑦ 혁신 원칙을 제시했다.

정부는 2024년 4월 '기업을 위한 인공지능 가이드라인(AI Guidelines for Businesses Ver. 1.0.)'을 발표했다. 위 가이드라인은 위 '인간 중심 인공지능의 사회적 원칙'을 바탕으로 제정되었다. 위 가이드라인 외에도 정부는 2023년 7월 '초중등 교육에서의 생성형 인공지능 활용에 관한 잠정 가이드라인(Provisional Guidelines for the Use of Generative AI in Primary and Secondary Education)'을 발표한 바 있다.

일본 정부는 2023년 12월 정보처리 추진기구(IPA) 내에 인공지능 안전연구소를 설립하였으며, 2024년 9월 '기업용 인공지능 가이드라인'과 인공지능 안전에 관한 국제 간행물을 바탕으로 인공지능 안전성 평가를 실시하기 위한 지침서인 '인공지능 안전성 평가 관점 가이드(Guideline to Evaluation Perspective on AI Safety)'를 발간했다.

인공지능의 개발 및 사용과 관련된 지식 재산권법에 대하여는, 2018년 저작권법이 개정되어 '저작물에 표현된 사상이나 감정을 향유할 목적이 없는 사용'과 '컴퓨터 데이터 처리 및 결과물 제공에 부수되는 사소한 사용'에 대한 새로운 저작권 예외 조항이 추가되었다. 이러한 새로운 예외 조항으로 인해 인공지능 개발에 있어 저작권이 있는 저작물의 특정 사용이 허용될 수 있다. 정부는 생성형 인공지능의 등장 이후 이러한 예외 조항이 정확히 어떻게 적용될 것인지에 대한 우려가 커지자, 2024년 1월경 전문가로 구성된 소위원회를 통해 '인공지능 저작권 이슈 개요 초안(Draft Overview of AI Copyright Issues)'을 발표해서 생성형 인공지능과 저작권을 둘러싼 이슈를 보다 명확하게 설명하고 있다. 정부는 2024년 5월에는 '인공지능 시대의 지식 재산권 연구반'을 통해 지식 재산권법과 인공지능 관련 이슈에 대한 논의의 중간 요약본을 발표하기도 하였다.

일본 정부는 2018년 6월에 '인공지능 및 데이터 사용에 대한 계약 가이드라인(Contract Guidelines for the Use of AI and Data)'을 발표하고 2018년 12월에 이를 업데이트했다. 구속력이 없는 이 가이드라인은 데이터 계약 또는 인공지능 사용 또는 개발을 위한 계약을 협상할 때 당사자가 알아야 할 기본 개념, 일반적으로 고려해야 할 문제 또는 요소, 모델

계약 등을 제시하고 있다.

가. 일본 총무성(MIC)·경제산업성(METI), 기업을 위한 인공지능 가이드라인(AI 事業者ガイドライン, AI Guidelines for Businesses Ver. 1.0., 2024. 4. 19.)

일본 총무성(MIC)과 경제산업성(METI)은 인공지능 개발자, 공급자 및 비즈니스 사용자에게 인공지능 거버넌스에 대한 통합된 지침 원칙을 제공하여 인공지능의 안전하고 보안적인 사용을 촉진하는 것을 목표로 가이드라인을 발표했다.

위 가이드라인이 인공지능을 통해 목표하는 사회는 ‘인간 중심 사회’이며, 이를 실현하기 위한 기본 철학이 ‘존엄성’, ‘다양성 및 포용성’, ‘지속가능성’이다. 그리고 인공지능 비즈니스 행위자가 존중해야 할 10가지 공통 지침 원칙으로, ① 인간 중심 인공지능, ② 안전, ③ 공정성, ④ 개인 정보 보호, ⑤ 사이버 보안, ⑥ 투명성, ⑦ 책임성, ⑧ 교육 및 리터러시, ⑨ 공정한 경쟁, ⑩ 혁신을 든다.

가이드라인은 히로시마 인공지능 프로세스에서의 정의를 차용하여 ‘첨단 인공지능’를 ‘최첨단 기초 모델과 생성형 인공지능 시스템을 포함한 가장 진보된 인공지능 시스템’이라고 정의한다. 그리고 첨단 인공지능 시스템에 관여하는 비즈니스 행위자는 위 공통 지침 원칙에 더하여 다음과 같은 사항을 준수해야 한다고 한다. ① 배포 및 시장 출시 전후를 포함하여 첨단 인공지능 시스템 개발 전반에 걸쳐 적절한 조치(레드팀 구성 등), ② 배포 및 시장 출시 후 취약점과 적절한 경우 사고 및 오용 패턴을 식별하고 완화, ③ 첨단 인공지능 시스템의 기능, 제한 사항 및 적절하고 부적절한 사용 영역을 공개적으로 보고, ④ 산업, 정부, 시민 사회, 학계 등 첨단 인공지능 시스템을 개발하는 조직 간에 책임 있는 정보 공유 및 사고 보고, ⑤ 위험 기반 접근 방식에 기반한 인공지능 거버넌스 및 위험 관리 정책을 개발, 구현 및 공개, ⑥ 인공지능 생애 주기 전반에 걸쳐 강력한 보안 관리 투자 및 구현, ⑦ 워터마킹 또는 신뢰할 수 있는 콘텐츠 인증 및 출처 메커니즘을 기술적으로 가능한 경우 개발 및 배포, ⑧ 사회적, 안전 및 보안 위협을 완화하기 위한 연구 및 효과적인 완화 조치에 대한 우선적 투자, ⑨ 기후 위기 등 세계 최대 과제를 해결하기 위한 첨단 인공지능 시스템 개발 우선, ⑩ 국제 기술 표준 및 모범 사례 개발에 참여 및 채택, ⑪ 개인 데이터 및 지식 재산에 대한 적절한 데이터 입력 조치 및 보호 구현, ⑫ 첨단 인공지능 시스템의 신뢰할 수 있고 책임감 있는 사용을 촉진하고 이에 기여하는 것(디지털 리터러시

·교육·인식 개선)이 그 추가 사항이다.

가이드라인에는 안전, 공정성, 개인 정보 보호, 사이버 보안, 투명성, 책임성 및 혁신 문
제와 관련하여 인공지능 개발자, 제공자 또는 비즈니스 사용자를 위한 별도의 지침이 있
다. 가이드라인에서 권장하는 거버넌스 접근 방식은, ① 인공지능 시스템 및 서비스와 관
련된 이점, 위험 및 외부 환경 분석 수행, ② 인공지능 거버넌스 목표 설정 및 인공지능 거
버넌스 목표 달성을 위한 인공지능 관리 시스템 설계 및 운영, ③ 인공지능 관리 시스템을
지속적으로 모니터링, 평가 및 지속적으로 개선하여 효과적으로 작동하는지 확인, ④ 규정
및 기타 사회 시스템의 변화를 포함한 외부 환경의 변화를 고려하여 환경 및 위험 분석을
재검토하고 필요에 따라 인공지능 거버넌스 목표 수정하는 것이 포함된다.

나. 일본 인공지능 안전연구소(AISI JAPAN), 인공지능 안전성 평가 관점 가이드(AIセーフ
ティに関する評価観点ガイド, Guide to Evaluation Perspectives on AI Safety
Version 1.01, 2024. 9. 25.)

일본 인공지능 안전연구소는 인공지능 안전성 평가 메커니즘과 원칙을 자세히 설명하는
가이드라인을 발표했다. 본 가이드라인은 인공지능 시스템 개발 및 제공업체와 관련자들
이 인공지능 안전성 평가를 수행할 때 도움을 주기 위한 것이다.

‘인공지능 안전성 평가’는 ‘인공지능 모델 및 시스템의 안전성 상태를 밝히고 안전성을
유지 또는 개선하는 것을 목표로 하는 포괄적인 관리 프로세스’이다. 그리고 기업을 위한
인공지능 가이드라인과 여러 국제 가이드라인에 의하면, ‘인간 중심’, ‘안전’, ‘공정성’, ‘개인
정보보호’, ‘보안 확보’, ‘투명성’이 인공지능 안전성의 핵심 요소이다. 이러한 핵심 요소를
기초로, 본 가이드라인은 안전성 평가를 수행하는 사람이 보다 구체적인 관점에서 평가
작업을 수행할 수 있도록 ‘평가 관점’ 10가지를 제시한다. ① 유해한 출력 제어, ② 허위 정
보, 거짓 정보 및 조작 방지, ③ 공정성 및 포용성, ④ 고위험 사용 및 의도하지 않은 사용
처리, ⑤ 개인 정보 보호, ⑥ 보안 확보, ⑦ 설명 가능성, ⑧ 견고성, ⑨ 데이터 품질, ⑩ 검
증 가능성 그것이다.

인공지능 안전성 평가의 주요 수행자는 인공지능 개발 및 제공 담당자이다. LLM 시스템
활용 흐름에는 개발, 제공, 사용의 세 가지 단계가 있으며, 시기적으로 인공지능 안전성 평
가는 각 단계에서 합리적이고 적절해야 하며, 반복적으로 수행되어야 한다. 인공지능 안전

성 평가 방법에는 ① 기술적 평가(도구 기반 평가, 레드 팀ing에 의한 평가, 벤치마크, 체크리스트, 밸런스 스코어카드 등 기타 기술적 평가 등)와 ② 관리적 평가(회사의 인공지능 안전에 대한 전반적인 정책 및 회사 내에 설정된 규칙 및 규정 등 평가)가 있다.

7. 우리나라

한국은 2024년 5월 'AI 서울 서밋'을 개최하고 '서울선언'과 그 부속서인 'AI 안전 과학에 대한 국제협력을 위한 서울 의향서'를 발표해 국제 공조와 협력의 기본원칙을 확인했다. 2024년 9월 26일 인공지능 산업 진흥을 위한 사항을 심의·조정하는 국가인공지능위원회가 수립되었고, 2024년 11월 27일 AI 안전연구소가 출범했다.

과학기술정보통신부는 2020년 12월 23일 '인공지능 윤리기준'을 발표하여 '인간성을 위한 인공지능'을 위해 사회구성원이 지켜야 할 3대 원칙(인간 존엄성, 사회 공공선, 기술의 합목적성)과 10대 요건(인권 보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성, 안전성, 투명성)을 제시하였다. 과학기술정보통신부는 2021년 11월 23일 '인공지능 윤리기준'을 실제 현장에 적용할 수 있는 방법을 구체화하여 '인공지능 윤리기준 실천을 위한 자율점검표(안)'를 발표하였다. 이후 과학기술정보통신부는 2023년 7월과 2024년 4월 일반, 공공사회, 의료, 자율주행, 채용, 스마트 치안, 생성인공지능 기반 서비스 분야에 관한 '신뢰할 수 있는 AI 개발 안내서'를 발표하였다. 그리고 한국정보통신기술협회(TTA)를 통해 2023년 12월 12일 인공지능 시스템 신뢰성 구성요소와 요구사항을 제시한 '인공지능 신뢰성 단체표준'을 제정하였다.

정부는 2023년 9월 25일 디지털 심화 시대에 맞는 국가적 차원의 기준과 원칙을 제시하고, 글로벌을 리드할 수 있는 보편적 디지털 질서 규범의 기본방향을 담은 현장인 '디지털 권리장전'을 공개하였다.

가. 금융 분야

금융 분야에서는 데이터 3법 개정으로 신용정보법 제36조의2에 자동화 평가에 대한 대응권 규정이 도입되었고, 금융위원회가 2021년 7월 8일 ① 기획·설계 단계 ② 개발 단계 ③ 평가·검증 단계 ④도입·운영·모니터링 단계별 고려사항을 제시한 '금융 AI 가이드라인'을 발표하였고, 금융혁신기획단이 2022년 8월 금융분야에서 인공지능이 가장 많이 활용

되는 5대 서비스별 'AI 개발·활용 안내서'를 체크리스트 형태로 제작하였다. 금융위원회는 2023년 4월경에는 'AI 기반 신용평가모형 검증체계'를 발표하여 개인신용평가체계 검증위원회로 하여금 활용하게 하였고, 인공지능 모델 개발단계별 보안 고려사항을 제시한 '금융 분야 보안 AI 가이드라인'도 발표하였다.

나. 지식 재산 분야

문화체육관광부는 2024년 1월 16일 '생성형 AI 저작권 안내서'를 마련하여 배포하였다. 인공지능 사업자에 대해서는 인공지능의 학습을 위한 저작물의 이용이 저작권법에 따른 '공정 이용'에 해당하는지 여부가 불명확하므로, 가급적 사전에 저작권자로부터 적법한 이용 권한을 확보하는 것이 바람직하다고 권고하였고, 저작권자에 대하여는 자신의 저작물이 인공지능 학습에 이용되는 것을 원하지 않을 경우 그에 반대하는 의사를 약관규정 등의 방식으로 명시하거나 이를 방지하기 위한 로봇배제표준 등 기술적인 조치를 취하는 것이 적절하다고 하였다. 인공지능 이용자에 대하여는 입력 데이터로 인해 타인의 저작권을 침해하거나 침해를 유도하지 않도록 유의할 것을 당부하였고, 인간의 창작적 개입이 있다고 볼 수 없는 인공지능 산출물에 대해서는 저작권 등록이 원칙적으로 불가하다고 하였다.

다. 개인정보 분야

2024년 3월 15일 시행된 개정 개인정보 보호법 제37조의2에 '자동화된 결정' 관련 의무가 신설되었다. 개인정보보호위원회는 2021년 6월 1일 '인공지능(AI) 자율점검표'를 발표하여 인공지능 설계, 개발·운영 과정에서 개인정보를 안전하게 처리하기 위하여 지켜야 할 개인정보 보호법 상 주요 의무·권장사항을 설명하였다. 개인정보보호위원회는 2023년 8월 3일 '인공지능 시대 안전한 개인정보 활용 정책방향'을 발표하여 인공지능 단계별(기획-데이터 수집-인공지능 학습-서비스) 개인정보 처리 기준을 구체화 하였다. 이후 개인정보 보호위원회는 비정형데이터에 대한 가명처리 기준을 담은 '가명정보 처리 가이드라인(2024. 2. 4.)'과 '인공지능(AI) 개발·서비스를 위한 공개된 개인정보 처리 안내서(2024. 7. 17.)'을 발표하였다.

라. 이용자 보호 분야

이용자 보호 분야에서 방송통신위원회는 2019년 11월 11일 이용자 중심의 지능정보사회를 실현하기 위한 '이용자 중심의 지능정보사회를 위한 원칙'을 발표하였고, 사람중심 서

비스, 투명성과 설명가능성, 차별금지 등의 내용을 담았다. 방송통신위원회는 2021년 6월 30일 디지털 플랫폼의 추천 서비스와 관련한 ‘인공지능 기반 미디어 추천 서비스 이용자 보호 기본원칙’을 공개하였고, 2023년 12월 18일 생성형 인공지능 이용자를 대상으로 생성형 인공지능 이용 시 염두에 두어야 할 사항을 정리한 ‘생성형 AI 윤리 가이드북’을 발간하였다.

마. 인권·교육·보안·공정거래 분야

국가인권위원회는 2022년 5월 11일 인공지능의 위험도 등급을 구분하고 그에 맞는 법제 마련을 촉구하는 ‘인공지능 개발과 활용에 관한 인권 가이드라인’을 발표하였다. 교육부도 2022년 8월 11일 교육분야 인공지능이 윤리적으로 개발되고 안전하게 활용될 수 있도록, 개발자와 교육당사자들이 함께 준수해야 할 ‘교육분야 인공지능 윤리원칙’을 발표하였다.

국가정보원은 2023년 6월 29일 챗GPT 관련 보안 문제를 사전에 예방하기 위해 ‘챗GPT 등 생성형 AI 활용 보안 가이드라인’을 공개하였다. 공정거래위원회도 2024년 하반기에 인공지능 시장에 대한 조사를 진행하였고 이번 시장 실태조사를 바탕으로 인공지능 시장의 혁신과 공정 경쟁을 지속할 수 있는 정책 방향에 관한 보고서를 마련할 예정이다.

바. 딥페이크 분야

2024년 1월 29일 시행된 개정 공직선거법에 따르면, 누구든지 선거일 전 90일부터 선거일까지는 딥페이크(deepfake) 영상 등을 선거운동을 위하여 제작·편집·유포·상영 또는 게시하여서는 안 되고, 그 외의 기간 중에는 딥페이크 영상 등이 가상의 정보라는 사실을 명확하게 인식할 수 있도록 표시하여야 한다. 그리고 2024년 9월 26일 성폭력처벌 특례법이 개정되어 허위영상물 편집·합성·가공·반포 행위의 처벌이 상향되었고, 소지·구입·저장·시청 행위 및 허위영상물 이용 협박·강요 행위를 처벌하는 규정이 신설되었으며, 같은 날 청소년성보호법도 개정되어 아동·청소년 성착취물 이용 협박·강요 행위를 처벌하는 규정이 신설되었다. 정보통신망 이용촉진 및 정보보호 등에 관한 법률(정보통신망법)도 2024년 12월 3일 개정되어 ‘인공지능 기술을 이용하여 사람의 얼굴·신체 또는 음성을 대상으로 한 촬영물·영상물 또는 음성물을 대상자의 의사에 반하여 편집·합성 또는 가공한 정보’라는 의미의 ‘합성영상등’ 개념이 도입되었고, 이에 정부는 관련 시책을 마련할 수 있게 되었다. 또한 이번 개정으로 방송통신위원회가 정보통신서비스 제공자나 게시판 관리

· 운영자에게 정보통신서비스 ‘성폭력범죄의 처벌 등에 관한 특례법’에 따른 편집물·합성물·가공물·복제물과 ‘아동·청소년의 성보호에 관한 법률’에 따른 아동·청소년성착취물에 대하여도 게시중단을 명령을 내릴 수 있는 근거 규정이 마련되었다.

사. 인공지능 기본법 입법 관련

‘국가정보화 기본법’은 2020년 6월경 ‘지능정보화 기본법’으로 전부 개정되어 4차 산업혁명 지원을 위한 범국가적 추진체계를 마련함으로써 데이터·인공지능 등 핵심기술 기반과 산업생태계를 강화하는 것을 목표로 하였다. 지능정보화 기본법 제2조 제4호는 지능정보기술의 하나로 ‘전자적 방법으로 학습·추론·판단 등을 구현하는 기술(가목)’ 등 또는 그 결합 및 활용 기술을 정의함으로써 ‘인공지능’ 기술을 포섭하고 있다.

제21대 국회에서는 고위험 인공지능 관련 사업자에 대한 고지의무 및 신뢰성·안전성 확보조치 의무와 생성형 인공지능에 대한 고지 및 표시 의무를 정한 ‘인공지능 발전과 신뢰 기반 조성 등에 관한 법률(대안)’이 과학기술정보방송통신위원회 전체회의에 계류되었다가 국회 회기 종료로 폐기되었다. 제22대 국회에서는 과학기술정보방송통신위원회가 전체회의에서 2024년 11월 26일 ‘인공지능 발전과 신뢰 기반 조성 등에 관한 기본법안(대안)’을 통과시켜 현재 해당 법안이 국회 법제사법위원회에 계류되어 있다. 향후 위 인공지능 기본법 관련 국회의 동향과 추후 제정될 시행령의 내용 등에 대한 지속적인 모니터링이 필요하다.

위 인공지능 기본법안에는 ‘사람의 생명, 신체의 안전 및 기본권에 중대한 영향을 미치거나 위험을 초래할 우려가 있는 인공지능 시스템으로서 아래 중 어느 하나의 영역에서 활용되는 인공지능 시스템’을 ‘고영향 인공지능’으로 규정하고 있고, ‘생성형 인공지능’를 ‘입력한 데이터(「데이터 산업진흥 및 이용촉진에 관한 기본법」 제2조 제1호에 따른 데이터)의 구조와 특성을 모방하여 글, 소리, 그림, 영상, 그 밖의 다양한 결과물을 생성하는 인공지능 시스템’으로 정의하고 있다. 고영향 인공지능 또는 생성형 인공지능을 이용한 제품 또는 서비스를 제공하는 인공지능 사업자는 이용자에 대한 고지 등 투명성 확보 의무가 있으며, 일정 기준 이상의 인공지능 시스템에 대하여는 인공지능 생애 주기에 걸친 위험 식별·평가·완화 등 안전성 확보 의무가 부과된다. 고영향 인공지능 관련 사업자에 대하여는 안전성·신뢰성을 확보하기 위한 조치에 대한 책무가 부과된다.

제3절 소 결

2022년 챗GPT가 대중에게 선보인 이후 생성형 인공지능 서비스가 급격히 확산하는 과정에서 기술에 대한 경험과 이해 부족에서 오는 여러 문제가 사회 논쟁거리가 되었다. 정보의 부정확성에 따른 문제뿐만 아니라 콘텐츠 생성을 인공지능에 의존하였을 때의 저작권 이슈 등이 대표적이었으며, 이와 함께 인공지능의 윤리적 이슈들도 해소되지 못한 채 오히려 심화하는 양상을 보였다. 이에 구성원들이 챗GPT 등을 적극적으로 받아들인 대학이나 공공성의 가치가 우선되는 공공기관에서 먼저 이용자들을 위한 활용 가이드라인이 개발되어 배포되었다.

앞서 살펴본 이 생성형 인공지능 활용 가이드라인들은 개발자와 사업자가 준수해야 할 윤리원칙에 준하는 지침 외에 이용자들이 알아야 할 생성형 인공지능에 대한 기본 이해와 활용 시 주의해야 할 점에 중점을 두고 있었다. 기존 기술도구나 인공지능서비스와는 달리, 생성형 인공지능 서비스는 이용자가 해당 서비스를 이용할 때 입력하는 명령어나 생성형 인공지능이 생성한 결과물에 대한 반응에 따라 최종 결과물이 달라진다. 즉, 모든 이용자가 같은 결과물을 얻는 것이 아니라 이용자-생성형 인공지능 간 상호작용에 따라 사회윤리에 어긋나지 않는 결과물을 얻거나 그렇지 않을 수 있고, 결과물의 정확성도 이용자의 입력 프롬프트(명령글/요구글)에 따라 좌우될 수 있다. 이에 초기 생성형 인공지능에 대한 가이드라인은 이용자 활용 가이드라인을 더 강조하게 된 것이다.

생성형 인공지능서비스의 기술 특성을 고려할 때, 활용단계에서 관리가 이루어지면 어느 정도 드러난 이슈들을 해소할 수 있지만, 또 다른 문제가 제기될 수도 있다. 이용자들은 생성형 인공지능의 기술 원리에 대한 이해가 부족한 상태에서 새로운 서비스를 접하였는데, 이용 책임도 온전히 이용자에게 부과되는 모양새가 될 수 있는 것이다. 실제로 서비스 제공자들의 이용정책이나 약관을 살펴보면, 현 시점에서 기술적 혹은 제도적으로 해결할 수 없는 이슈들 대부분이 서비스 활용단계에서 이용자 책임 범위로 되어 있는 경우가 많았다. 이는 신기술 기반 서비스가 막 도입되는 시기에 가장 정보력이 낮은 이용자에게 책임이 쏠리게 되는 것은 이용자 보호 측면에서 문제가 될 수 있다.

다음으로 살펴본, 국내의 인공지능 관련 규제 동향을 보면, 개발사와 서비스 제공자의 책임 수준을 자율 규제뿐만 아니라 법적 규제로도 통제하려는 경향을 확인할 수 있었다.

아직은 법적 규제 범위가 제한적이고, 민간 사업자에 대한 관리·통제는 자율규범 형식이 더 많으나 인공지능 생성 콘텐츠를 알릴 의무 등의 투명성 등 실천 가능한 법제 확립으로 논의의 방향이 진행되고 있었다.

특히 글로벌 동향을 종합하면 다음과 같이 유형을 구분해볼 수 있었다. 첫째, EU와 같이 포괄적인 인공지능 기본법과 관련 거버넌스 법령(경성법)을 토대로 인공지능에 대한 실질적이고 유효한 규제를 하는 유형, 둘째 미국과 같이 포괄적인 인공지능 기본법이 부재하는 등으로 경성법의 영향력이 약하고 이를 시장의 자율에 맡기지만, 정부가 시장의 건전한 관행 형성을 위해 강력한 중앙집권적 지침(연성법)을 시행하는 유형, 셋째, 영국과 같이 포괄적인 인공지능 기본법은 없으나 인공지능 이슈를 기존 법령을 통해 적극 포섭하거나 필요한 경우 기존 법령을 개정하며, 법령 해석 가이드라인 또는 건전한 시장 관행 형성을 위해 소관 규제기관이 가이드라인을 적극 발간하는 유형으로 구분할 수 있다.

우리나라의 경우 방송통신위원회 등 소관 부처가 인공지능 이슈에 대해 시의적절하게 가이드라인을 개발, 발표하여 이용자 보호에 앞장서고 있으며, 다른 한편에서는 인공지능 기본법을 제정하기 위한 노력이 계속되고 있다. 현재 발의된 인공지능 기본법은 기술 개발 및 산업 발전에 대한 진흥 정책뿐만 아니라 어느 정도 실효성 있는 위험관리 체계 정립 부분도 포함되어 있다. 따라서 우리나라의 유형은 현시점에서는 영국에 가까운 유형이지만, 향후 인공지능 기본법이 제정되고 기본법에 제재 규정을 포함한다면 형식에서는 EU와 같은 경성법 중심의 유형이 될 수도 있다.

기본법이 제정된다고 하더라도 한동안은 자율규범과 병행해야할 것이다. 법 제정 노력과는 별도로 우리나라는 인공지능 기술 연구개발이나 산업 발전에 저해되지 않으면서 위험을 효율적으로 관리할 방안을 모색하고 있기 때문이다. 이에 현시점에서 우리나라의 인공지능 규제 방향은 시장의 자율규범에 맡기면서도 정부가 시장의 건전한 관행 형성을 위해 지침을 개발하고 시행하는 미국의 유형에 가깝게 가는 것이 적절하다고 본다.

제3장 생성형 인공지능 서비스 이용자 보호 가이드라인 개발 과정

이번 장에서는 생성형 인공지능 서비스 이용자 보호 가이드라인 개발 과정을 정리하였다. 먼저, 2024년 2월 연구반을 구성하고 키포트를 하면서 위원들 간 공유한 가이드라인 개발의 필요성과 목적을 밝힌다. 다음으로 연구반 구성원 소개와 회의 내용을 정리하였다. 회의는 연구반 구성원 모두가 참여하는 총 6회의 전체회의를 중심으로 정리하였으며, 내용 작성을 위해 따로 꾸린 실무반의 회의는 생략하였다. 실무반의 작업과 회의는 주로 전체회의에서 나온 내용을 문서화하고 수정, 보완하는 데 집중하였기 때문이다. 셋째, 가이드라인 개발 과정에서 가이드라인의 수범자이자 직접적 이해당사자인 사업자들의 의견을 수렴하였는데, 그 내용을 간략히 정리하였다. 사업자의 의견은 먼저 개발 초기 시점에 서면 조사를 통해 사업자가 이용자 보호를 위해 어떤 조치를 하고 있는지를 사전에 확인하였고, 최초의 초안이 나온 이후 국내 사업자와 대면 면담 형식으로 의견을 수렴하였다. MS, OpenAI, 구글 등의 글로벌 기업의 경우, 본사의 검토가 필요하여 직접 대면 대신 서면으로 의견을 수렴하였으며, 한글 문서를 다시 영문화하여 본사에서 검토하고 다시 한글화하여 전달되는 과정을 거쳤다. 마지막으로, 사업자 의견을 반영한 수정본을 다시 연구반 위원들이 검토하고 방송통신위원회의 최종 검토를 거쳤다. 그 결과물은 이 장의 4절에 실었다.

제1절 가이드라인 개발의 필요성 및 목적

1. 이용자 보호의 의미와 생성형 인공지능 시스템의 특수성

어떠한 제품이나 서비스를 막론하고 주어진 환경 아래에서 이용자가 해당 제품 및 서비스를 신뢰하고 안전하게 사용하거나 영위할 수 있도록 보장하는 일은 중요하다. 이는 제품이나 서비스를 제공하는 기업의 입장에서는 사업을 유지하고 확장성을 가지기 위한 조

건이 되지만, 법·제도적인 측면에서는 일반 시민에 해당하는 이용자를 법질서가 보호한다는 점에서 의미를 가진다. 따라서 이용자 보호는 중요한 과제이며, 일반적인 소비자보호 법제뿐만 아니라 많은 영역에서 특유의 이용자 보호 제도들이 발달하고 있다. 방송통신의 영역에서는 전기통신사업법을 비롯한 관련 법제들이 이용자 보호에 대해 규율하고 있다(김태오, 2009; 박중수, 2013; 최경진 외, 2015). 주로 통신서비스를 매개로 이루어지는 인공지능 시스템에 대한 이용자 보호 역시 인공지능에 대한 일반법이 존재하고 있는 우리나라의 법적 상황 속에서는 전기통신사업법을 근거로 한 이용자 보호의 거버넌스에 포함될 수 있을 것이다. 특히나 방송통신정책이 더 이상 네트워크의 관리에 머무르지 않고, 통신서비스 안에서의 콘텐츠 향유나 이용자의 행태에 관심을 두게 되는 변화를 주목할 필요가 있다. 이에 따라 그 안에서 서비스 제공의 수단으로 활용하게 되는 인공지능 시스템과 같은 범용기술에 대한 규율이 이용자 보호 차원에서 더욱 중요해진 것이다(이원태 외, 2018).

일반적인 인공지능 시스템뿐만 아니라 생성형 인공지능을 기반으로 한 시스템 역시 이용자 보호 이슈가 제기된다. 생성형 인공지능의 대두는 사실 통신 기술, 클라우드 및 하드웨어의 발달에 기인한 것으로 추론할 수 있다. 생성형 인공지능의 기반인 대규모 언어모델을 훈련시키기 위해 필요한 물적 부담이 기술의 발달로 인해 대폭 감소한 것이다. 따라서 대규모의 데이터 처리와 학습 모델의 축적이 생성형 인공지능을 가능케 하였고, 기존에는 상상 속에서만 머물렀고 현실화되리라고 생각하기 어려웠던 이른바 인공일반지능(artificial general intelligence: AGI)으로의 발전 가능성 역시 운위되고 있는 상황이다(이화민, 2023; Susnjak et al., 2024). 네트워크 기술의 발전에 따라 방송통신의 이용자들은 과거보다 개별화되고 능동적인 태도를 보여주고 있다(이원태 외, 2018). 그러나 생성형 인공지능의 경우 이용자의 개별성은 더욱 강화될 수 있지만, 이용자의 능동성을 강화한다고 말할 수 있을지는 의문이다. 우선 신기술의 도입 초기 단계로서 이용자의 경험과 역량이 성숙되지 않았다는 문제가 있다. 또, 이용자의 대응에 생성형 인공지능이 반응하는 상호작용이 반복되면서 오히려 이용자의 능동성이 감소하는 효과가 있다.

이용자 보호의 관점에서 이용자가 해당 서비스를 지속 가능하게 이용할 수 있도록 지원하는 것이 중요하다. 이를 위해 생성형 인공지능이 야기할 수 있는 리스크를 적절히 제어할 수 있는 방안이 필요하며, 이에 생성형 인공지능이 이용자에게 미치는 영향과 이에 대응하는 법질서 차원의 수단들을 고려하여 생성형 인공지능 서비스 이용자 보호 가이드라

인(이하 ‘가이드라인’)을 개발하고자 하였다.

다음은 기존의 인공지능 관련 규범들과 윤리적 논의들을 통해 생성형 인공지능에 적용 가능한 법리들을 살펴본 것이다. 이미 2장에서 살펴본 내용과 겹치는 부분도 있으나 개발 과정을 충실히 담는다는 차원에서 다시 정리하였다. 이 법리들은 가이드라인을 개발할 때 고려할 주요 기준이었다. 즉, 생성형 인공지능 시스템이 야기할 수 있는 리스크를 분석하고 각각에 대한 대응방안을 도출하는 한편, 종합적으로 우리 법질서가 생성형 인공지능에 대해 어떠한 태도를 취해야 할지 고찰하는 기준으로 활용하였다.

2. 제도적 대응 사례를 통해 본 가이드라인 개발의 필요

생성형 인공지능에 대해서는 아직 제도적인 대응이 불충분한 상황이지만, 기존의 인공지능 관련 규율 논의들의 추이를 통해 생성형 인공지능에 대한 규율 방향성을 예측할 수 있다. 인공지능 관련 규율은 크게 두 가지의 방향성으로 구분할 수 있는데, 첫 번째 시각은 인공지능을 경제 발전의 동력으로 인식하고, 산업진흥의 장애가 되는 리스크들을 제거하는 방향으로 법·제도를 개발하는 것이다. 두 번째 시각은 인공지능으로 인한 리스크가 사회에 미치는 해악 방지에 초점을 두고 윤리적 접근에서 법적 규제를 정립하는 것이다. 생성형 인공지능에 대해서도 이와 같은 방향성 구분을 적용해 살펴볼 수 있다. 다음의 글로벌 법제 동향 고찰에서 이 두 시각이 드러난다.

가. 유럽연합의 윤리적 접근에서 법적 규제로 변화

유럽연합의 인공지능 관련 규범은 융커(Jean-Claude Juncker) 집행위원장 시기의 윤리적 접근에서, 라이엔(Ursula von der Leyen) 집행위원장 시기의 법적 규제로 변화한 것을 발견할 수 있었다. 2019년 집행위원회가 전문가들을 초청하여 구성한 “인공지능 고위급 전문가그룹(High-Level Expert Group on AI: HLEG)”은 “신뢰할 수 있는 인공지능을 위한 윤리 가이드라인(Ethics guidelines for trustworthy AI)”을 발표하였다.¹⁾ 여기에는 윤리적 원칙

1) 해당 가이드라인의 전문은 다음 사이트에서 확인할 수 있다. Ethics guidelines for trustworthy AI | Shaping Europe's digital future
<<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>>(2024. 7. 22. 최종 접속)

과 관련된 가치를 비롯하여 인공지능에 대한 신뢰를 위해 필요한 7가지 핵심 요건과 신뢰성 있는 인공지능 평가 목록을 담고 있다.

2019년 12월에 취임한 라이엔 집행위원장은 디지털 분야에서의 적극적인 법적 규율 의지를 천명하고, 법적 규율의 준비 작업으로서 우선 「인공지능 백서(White Paper on Artificial Intelligence)」(2020)를 발간하였다. 이 백서에서 집행위원회는 EU 시민의 가치와 권리 존중, 윤리적이고 신뢰할 수 있는 인공지능의 개발 및 활용, 인공지능 분야 유럽의 혁신 역량 증진을 위한 정책 방향을 제시하고 있다. 또한 유럽 내 단일 인공지능 생태계 조성을 위해 ① 우수한 생태계(ecosystem of excellence) 조성을 위한 프레임워크, ② 신뢰의 생태계 조성을 위한 규제 프레임워크를 제시하고 있다. 이 규제 프레임워크에서는 학습데이터에 대한 관리에서 시스템 자체의 견고성과 안정성 및 정확성의 보장 및 인간의 감시 및 개입 보장에 이르기까지의 규제 체계를 제안한다. 특히, 인공지능을 활용한 원격 생체 인식 기술 적용과 같이 리스크가 큰 시스템의 사용에 대해서는 엄격하게 제한할 것을 제안하였다. 유럽연합 데이터 보호 규정(GDPR)에 따라 생체(biometric) 정보의 처리는 원칙적으로 금지되며, 예외적으로 법률의 근거에 따라 상당한 공익상(substantial public interest)의 이유가 있는 경우에 처리를 허용할 수 있으나, 인공지능 시스템을 통한 활용은 더 엄격하게 규제하고자 했던 것이다. 이러한 규제 프레임워크의 적용 대상으로는 기본적으로 사전 적합성 평가를 통해 판명된 고리스크 인공지능 애플리케이션을 제시하고 있으나, 그 밖의 애플리케이션에 대해서도 자발적 라벨링(labeling)을 유도한다는 방침을 밝히고 있다.

유럽연합은 이 백서의 내용에 기반한 회원국 및 이해관계자들의 의견 수렴을 거쳐 이른바 “인공지능법(AI Act)”을 입법한다(라기원, 2024). 2021년에 발의되어 2024년에 의회와 평의회의 의결을 마친 인공지능법(AI Act)은 인공지능 개발 및 사용을 위한 공통된 조건을 만들어 유럽 단일 시장의 기능을 보장하기 위한 목적에서 인공지능에 대한 리스크 기반의 규제를 골자로 하는 규제 입법이다. ① 안전하고 기존 EU 법률을 존중하며, ② 투자를 촉진하기 위해 법적 안정성을 보장하고, ③ 유럽법 차원의 기본권과 및 안전 요건 보장을 위해 규제를 효과적으로 집행할 수 있는 구조를 조성하고, ④ 합법적이고 안전하며 신뢰할 수 있는 인공지능 애플리케이션을 위한 단일 시장의 성장을 촉진할 것을 목적으로 하는 제2차법에 해당하는 규정이다. 그 핵심 내용은 리스크 기반의 통제라고 할 수 있다. 즉, 인

공지능 시스템에 대해 리스크별로 분류한 후, 리스크의 정도에 따라 규제 정도를 달리 하는 방식을 달리하는 것이다. 허용할 수 없는 정도의 리스크(unacceptable risk)를 안고 있는 인공지능 시스템(예: 인간을 수단으로 삼는 인공지능 시스템이나 예측 기반 경찰 활동)은 전면적으로 이용을 금지하고, 높은 리스크(high risk)를 가진 시스템(예: 신용평가, 건강 및 생명보험, 투표 집계 시스템, 의료 분야에서의 환자 분류 등)은 엄격한 규제를 통해 제한된 조건 아래에서만 활용이 가능하도록 하고, 활용 이후 사후 규제 역시 강화하는 것이다. 반면 제한된 리스크(limited risk)만 가지고 있는 시스템에 대해서는 사고 발생 시 사후 대처가 가능하도록 하는 투명성 의무를 부과하고, 최소의 리스크를 가졌거나 리스크가 없는 것으로 평가되는(minimal or no risk) 인공지능 시스템에 대해서는 자유로운 사용을 보장한다. 또한 범용 인공지능에 대해서는 그 모델의 제공자에 대해 학습 과정 등에 대한 기술적 문서를 작성할 의무를 포함한 특수한 의무를 부여하고 있다. 인공지능법이 예정하는 인공지능 규율 거버넌스는 유럽 인공지능 위원회(European Artificial Intelligence Board), 독립한 전문가 패널 및 회원국 내국의 경쟁당국으로 삼분된다. 아울러 2024년 2월에 이미 집행위원회 산하에 유럽 인공지능청(European AI Office)을 설치하여 규칙의 집행과 적용을 위한 준비를 하고 있다.

유럽연합은 여전히 인공지능을 장래의 성장 동력으로 삼으려는 정책 기초를 유지하면서도, 더욱 안전하고 더 신뢰할 수 있는 인공지능 구현을 위해 향후에도 그 오용으로 인한 리스크에 적극 대처하겠다는 입장을 보였다. 기존의 인공지능법 뿐만 아니라 향후 산업에서의 인공지능 활용에 관한 이니셔티브를 담은 입법 움직임이 있을 것으로 보이며, 인공지능 연구 이사회(European AI Research Council) 등의 거버넌스도 집행위원회 차원에서 준비하고 있다고 공표하였다(Leyen, 2024.7.18.). 명시적인 언급은 없지만 생성형 인공지능에 대한 대응 역시 인공지능 기술 발전에 발맞추어 더 적극적으로 이루어질 것으로 보인다.

나. 주요 국제기구 및 단체들의 연성규범을 통한 접근

주요 국제기구들과 단체들 역시 구속력이 있는 법규가 아닌 권고안 등의 연성규범을 통해 국제사회 혹은 회원국에게 인공지능 정책의 방향성을 제시하고 있다.

경제협력개발기구(OECD)는 디지털 변혁의 시대에서 지속가능한 발전과 인공지능 활용을 위해 OECD 회원국 간의 논의를 거쳐 2019년 5월에 「OECD 인공지능 이사회 권고」(OECD Recommendation of the Council on Artificial Intelligence)를 발표하였다(OECD,

2019). 향후 회원국들의 경제개발정책의 새로운 성장 동력으로서 인공지능의 중요성을 인정하고, 역기능을 최소화한 채로 인공지능을 최대한 활용하기 위한 전략을 제시하는 내용이다. OECD는 인공지능을 통한 포괄적 성장, 지속가능한 발전 및 삶의 질(웰빙)(inclusive growth, sustainable development and well-being)을 달성하기 위해 몇 가지 개별 원칙들을 제시하고 있다. 인간 중심적 가치와 공정성(human-centered values and fairness), 투명성과 설명가능성(transparency and explainability), 견고성, 보안 및 안전성(robustness, security and safety) 및 책임성(accountability)이 이에 해당한다. OECD는 지난 2024년 5월, 생성형 인공지능을 비롯한 기술의 발전 사항들을 반영하여 본 권고안의 내용을 개정하여 발표하였다. 여전히 신뢰할 수 있는 인공지능을 위한 책임 있는 관리 임무(responsible stewardship)를 요구한다는 기초를 유지하면서, 권고의 대상을 “인공지능 시스템”으로 변경하고 있다. 또한 생성형 인공지능의 맥락에서는 정보의 무결성을 유지할 것을 요구하고 있다.²⁾

전기전자기술자협회(IEEE)는 지난 2019년 9월, 산하 프로그램 중 하나인 “자율적이고 지능적인 시스템에 대한 국제 윤리 이니셔티브(The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems)”를 통해 인공지능 윤리에 대해 기술자 및 기술 개발 기업이 준수해야 하는 사항을 담은 “윤리적 설계를 위한 문서(Ethical Aligned Design document)”를 발간했다. A/IS가 국제적으로 승인된 인권을 존중하고 고양시키며 보호하기 위한 목적으로 개발 및 운영되어야 한다는 인간 권리에 대한 내용을 포함하여, 인간 생활 웰빙의 개선, 데이터 에이전시, 효율성, 투명성, 책임성, 오남용의 인식을 설계의 목표로 삼아야 한다는 점을 제시하고 있다. 또한 개발자들이 A/IS³⁾의 안전하고 효율적인 운용을 위해 필요한 지식과 기술을 명시할 의무도 부여하고 있다.

국제연합교육과학문화기구(United Nations Educational Scientific and Cultural Organization: UNESCO) 역시 국제적인 교육, 과학 및 문화의 가치를 인공지능의 개발과 사용에 대해 관

2) 관련 내용에 대해서는 OECD Legal Instruments

<https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#backgroundInformation>(2024. 7. 22. 최종 접속)

3) IEEE는 인공지능(Artificial Intelligence)이라는 용어 대신 자동화/지능화 시스템(Autonomous and Intelligent System)이라는 용어를 사용한다.

철할 목적으로 권고안(The Recommendation on the Ethics of Artificial Intelligence)을 2020년에 발표한다(이상욱, 2021). 유네스코가 추구하는 가치에 맞추어 인공지능 윤리를 위한 가치로 인간의 존엄성, 인권 및 근본적 자유, 소외된 사람이 없을 것, 조화로운 삶, 신뢰가능성 및 환경보호를 제시하고 있다. 이를 달성하기 위한 실천 규범은 두 가지 범주로 제시한다. 첫 번째 범주는 인간과 인공지능 간의 관계로서, 인간과 인간의 변형, 비례성, 인간의 관리감독 및 결정, 지속가능성, 다양성과 포용성, 개인정보 보호, 인식과 교육 및 다중 이해관계자 및 적응형 거버넌스를 세부 원칙으로 제시한다. 두 번째 범주는 인공지능 시스템 자체에 대한 내용으로 공정성, 투명성 및 설명가능성, 안전과 보안 및 책임과 책무성 등을 강조한다.

다. 미국의 법안과 정부입법

미국 정부와 의회 역시 다양한 형태의 인공지능 관련 규범을 입법하거나 입법하기 위해 시도하고 있다. 연성규범 및 행정입법 형태로는 몇몇 규범들이 발효되었으나, 유럽연합과 달리 법률 차원의 규범은 아직 제정되지 않은 상태이다. 본래 미국 정부의 기조는 내국 기업들의 우월한 인공지능 관련 기술을 바탕으로 인공지능의 활용성을 최대한 확장하는 것이었으나, 최근 생성형 인공지능 등 기술의 확대에 의해 인공지능 시스템의 리스크 제어에도 어느 정도 시선을 두고 있는 것으로 보인다.

우선 바이든(Joe Biden) 정부 초기에 “연방정부의 신뢰할 수 있는 인공지능 사용을 도모하기 위한 행정명령(Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government)”을 발한 것을 본격적인 인공지능 규범의 시작으로 볼 수 있다. 이 행정명령은 연방정부 내에서의 인공지능 사용에 대한 통제를 목적으로 하는 것으로, 민간 영역에서의 인공지능 사용을 규율하는 것은 아니다. 이 명령에서는 인공지능 사용의 원칙을 a) 합법적이면서 미국의 가치를 존중하는 사용, b) 목적 및 성과 중심의 사용, c) 정확하고, 신뢰할 수 있으며 효과적인 사용, d) 안전하고 회복 가능한 사용(safe, secure, and resilient), e) 합리적인(understandable) 사용, f) 책임성 있고 추적 가능한 사용(responsible and traceable), g) 정기적인 모니터링, h) 투명성, i) 결과에 대한 책임(accountable) 등으로 제시하고 있다. 또한 이 원칙들이 실효성을 가질 수 있도록 하는 감독 기구 설치, 사용 사례(use case)의 개발, 기관 간 협력, 외부(산업 및 학계) 전문가의 참여 등을 제시하고 있다. 정부 내의 인공지능 시스템 사용에 적용되는 규범이지만, 공공 입

찰을 통해 참여하는 민간 사업자들에게도 명령을 적용함으로써 일부 우회적인 규제가 이루어진다.

의회에서의 인공지능 관련 입법 시도로는 와이든(Ron Wyden)과 부커(Cory Booker)(이상 상원) 및 클라크(Yvette Clarke, 하원) 의원이 발의한 알고리즘 책임법안(Algorithmic Accountability Act of 2022)을 들 수 있다. 자동화된 의사결정 시스템에 대한 투명성과 책임성 확보를 목적으로, 기업이 이용자와 규제기관에 ‘자동화된 의사결정 프로세스’가 어떻게 활용되는지를 인식할 수 있도록 관련 정보를 제공함으로써 투명성을 확보하고, 기업이 그 의사결정의 영향평가를 지속적으로 수행하도록 하는 것이 골자이다. 적용 대상은 자동화된 의사결정시스템(automated decision system; ADS)과 증강된 중요 의사결정 프로세스(augmented critical decision process; ACDP)로서 사람이 최종적인 의사결정을 하는 경우에도 그 의사결정이 인공지능·알고리즘 활용 결과를 근거로 할 때는 적용 대상이 된다.

“인공지능 권리장전 청사진(Blueprint for an AI Bill of Rights, 2022)”은 백악관 과학기술 정책국(OSTP)이 인공지능 시스템 설계·이용·배포 과정에서의 시민 권리 보호를 위한 지침서로 발간한 것이다. 이 권리장전에서는 시민 권리 및 민주적 가치 보호를 위한 5대 원칙을 밝히고 있다. 안전하고 효과적인 시스템(safe and effective systems), 알고리즘 차별 금지(algorithmic discrimination protections), 개인정보보호(data privacy), 고지 및 설명(notice and explanation), 인간 대안, 검토 사항, 비상 시 대책(human alternatives, consideration, and fallback) 등이 그 내용이다. 아울러 5대 원칙의 실행을 위한 별도 핸드북도 발간하고 있다. 이 핸드북에서는 산업, 정부, 지자체 및 기타 관계자들이 위의 보호 원칙을 정책, 관행 및 기술 프로세스에 관철하기 위해 취해야 할 예시를 제공한다.

생성형 인공지능이 본격적으로 대두하기 시작한 2023년에 연방정부는 다시 “안전하고 신뢰할 수 있는 인공지능의 개발과 사용을 위한 행정명령(Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence)”을 발령한다. 앞선 2020년의 행정명령과는 달리 공공부문에서의 인공지능 시스템 사용뿐만 아니라, 국가 안보, 건강, 안전을 위협하는 인공지능 기술 개발과 이용을 규제하겠다는 의도를 담고 있다. 주요 내용으로는 인공지능 개발기업의 안전성 평가 의무화, 인공지능 도구의 안전성 표준 수립, 가짜뉴스 및 딥페이크 검증을 위한 콘텐츠 인증 표준 수립 및 개인정보 보호 등이 있다.

라. 영국의 표준화 정책

영국은 어느 영역에서건 경성의 규범을 통한 규율보다는 표준화 또는 거버넌스를 통한 접근을 선호해 왔다. 인공지능과 관련해서도 2021년에 발간한 “국가 알고리즘 투명성 표준(algorithmic transparency data standard: ATRS)”이 일정한 규범적 역할을 수행한다. 이 표준은 데이터윤리혁신센터(CDEI)의 권고에 따른 후속조치로서, 공공부문의 조직이 알고리즘(algorithmic tool)을 사용하는 경우 투명성 의무를 부여하여 투명성 보고서를 발간하게 하는 것이다. 투명성 보고서는 3단계로 나뉘어 의무화되어 있다. 먼저 제1단계에서는 표준의 적용 대상인지의 여부에 대해 검증하고 공개 대상 정보를 선정한다(원칙적으로 전부 공개하므로 사실상 예외 선정). 제2단계는 알고리즘 투명성 보고서의 작성의 단계이다. 알고리즘 기능과 의사결정 절차, 알고리즘의 개발과 적용에 대한 정보, 결정 이유에 대한 설명, 기술적 특수성과 데이터, 리스크 및 영향분석 결과를 보고서에 신는다. 이러한 보고서를 공개하는 것이 제3단계의 과업이다.

마. 생성형 인공지능 이용자 보호 의제에의 시사점

위에서 제시한 인공지능 관련 규범들을 생성형 인공지능 시스템의 이용자 보호 정책에 참고하기 위해서는 우선 각각의 법제 및 윤리원칙의 목적과 특징에 대한 분석이 필요하다. 주요국 및 국제기구들은 인공지능과 관련하여 자국 혹은 해당 단체의 입장을 반영하고자 노력하였다. 따라서 위에서 언급한 대로 정도의 차이는 있지만, 인공지능 관련 규율의 방향성을 어떻게 선택했는지에 따라 규율의 내용이 상이해지는 것이다. 선택한 방향성에 따라 발표 주체별로 강조점이 달라지며, 규율의 범위와 시각 역시 차이를 보이게 된다.

생성형 인공지능 시스템에 대한 이용자 보호 정책을 연구한다는 차원에서 위 내용들을 어떻게 활용할 수 있을지 고찰이 필요하다. 우선 기존의 국내외 인공지능 규범의 내용을 그대로 수용하기보다는 생성형 인공지능에 대한 것이라는 점, 이용자 보호의 목적성이 있다는 점 및 인공지능 관련 산업 생태계와 문화에 차이가 있다는 점을 고려할 필요가 있다. 이에 법적·사실적 상황에 부합하는 원칙과 정책으로 해외의 사례를 번안할 필요가 있다. 또한 규범의 내용뿐만 아니라, 규범을 효과적으로 이행하기 위하여 필요한 구조와 실행 방식에 대한 탐구가 필요하다. 인공지능에 대한 윤리적 원칙을 포함한 많은 규범이 실행 구조를 통해 뒷받침되지 않는다면, 공허한 구호에 불과할 것이기 때문이다. 인공지능 시스

템 기반 애플리케이션의 제공자뿐만 아니라, 다양한 당사자들이 공동으로 역할을 수행할 수 있는 법제도적 방안도 연구할 필요가 있다. 예컨대 최종 이용자의 역할과 의무에 대해서도 어느 정도 고려해야 할 것이다. 관련 거버넌스에서 정책 집행자가 참여하는 정도와 방식에 대한 논의도 이루어져야 할 것이다.

제2절 가이드라인 개발을 위한 연구반 구성 및 운영

이번 절에서는 가이드라인 개발을 위해 구성한 연구반을 소개하고, 주요 회의의 내용을 정리하였다. 생성형 인공지능 서비스 이용자 보호 자율규제 가이드라인 개발을 위한 연구반은 미디어·법·기술·소비자 등 다양한 분야의 전문가를 비롯하여 관련 협회를 대표하는 영역에서 총 9인의 위원들로 구성되었다. 연구반 위원들의 명단은 다음과 같다.

〈표 3-1〉 생성형 인공지능 서비스 이용자 보호 가이드라인 연구반 위원 명단

분야	성명	소속 및 직위
미디어	이재신	중앙대학교 미디어커뮤니케이션학부 교수
법	윤종수	법무법인 광장 변호사
	선지원	한양대학교 법학전문대학원 교수
기술	김현철	고려대 컴퓨터학과 교수
	유지연	상명대 휴먼지능정보공학과 교수
	이대호	성균관대 인공지능융합학과 교수
소비자	김경자	가톨릭대 소비자학과 교수
시민단체	정지연	한국소비자연맹 사무총장
산업	김영규	한국인터넷기업협회 정책실장
공공	성재식	방송통신위원회 인공지능이용자보호대응팀 팀장
	조성은	KISDI 연구위원

2024년에는 연구반 전체회의를 총 6회 개최하였으며, 가이드라인 관련 국내외 규범을 고찰하고, 가이드라인 모델·내용 수준을 합의하였으며, 생성형 인공지능 서비스 이용자 보호를 위한 주요 쟁점 사항 및 가이드라인 초안에 대해 논의하였다.

2024년 한 해 동안 연구반 전체회의를 통해 진행된 주요 논의사항은 다음과 같이 요약되며, 여섯 차례 전체회의에서 이루어진 상세한 논의 내용은 다음에서 소개한다.

〈표 3－2〉 2024년 생성형 인공지능 서비스 이용자 보호 가이드라인 연구반 개최 현황

회차	개최 일자	주요 논의사항
1차	2024. 2. 29.	• 연구반 킷오프 • 주요 생성형 인공지능 서비스의 주제별* 이용자 보호 조치 현황 (이용자 고지 내용 등) 공유 * 책임소재 및 이용자 의무, 정보 정확성, 학습 데이터 활용, 권리 침해 신고 절차 등 • 생성형 인공지능 서비스 위험 유형 및 쟁점 논의
2차	2024. 3. 28.	• 국내 기발표된 생성형 인공지능 서비스 가이드라인 검토 • 해외 인공지능 규범 동향 검토 • ‘이용자’ 개념 및 범위 가이드라인 대상 범위 논의
3차	2024. 4. 19.	• 문제·이슈별 해결을 위한 원칙·기준, 구체적 실천 방안, 실제 사례 및 현황 등에 개별 의견 수렴 후 토의
4차	2024. 6. 18.	• 가이드라인 초안에 대한 전체 토의
5차	2024. 8. 16.	• 사업자 의견 공유 • 가이드라인 수정(안) 공유 및 향후 방향 토의
6차	2024. 10. 18.	• 가이드라인 최종 수정(안)에 대한 전체 토의

자료: 연구진 작성

1. 제1차 가이드라인 연구반 전체회의(킷오프)

연구반 제1차 전체회의는 2월 29일(목) 오후 2시 토즈스터디센터 양재점에서 대면 회의로 진행되었다. 이날 생성형 인공지능 서비스 이용자 보호를 위한 자율규제 가이드라인 마련을 위한 연구반 킷오프를 주제로 방송통신위원회, 정보통신정책연구원, 학계·산업계·시민단체 소속 위원 등 12명이 연구반 위원으로 참석하였다. 주요 내용은 연구반 구성원 소개와 운영 목적 및 계획 공유, 그리고 향후 개발할 가이드라인의 성격·구성(안), 사전 작업으로 진행할 사업자 의견 청취를 위한 문항 논의 및 향후 일정 등을 논의하였다. 주요 논의 내용은 다음과 같다.

먼저, 이용자에 대한 개념과 범위를 설정하고자 하였다. 생성형 인공지능 서비스는 그

유형별(챗봇, 콘텐츠 등)로 이용자 양상이 다르게 나타나므로, 일괄적으로 이용자 범위를 ‘엔드유저(end user)’로만 설정하기는 어렵다는 의견이 있었다. 이어 영역별 접근이 필요하다는 의견이 나왔으며, 그 이유는 생성형 인공지능 서비스는 개발자 중심으로 통제해 왔던 일반적인 인공지능 서비스와는 다르게 이용자가 어떻게 사용할지에 대해 예측 및 통제하기가 어렵다는 점에서 큰 차이가 있기 때문이었다. 전체적으로 보호 대상을 ‘서비스 이용자’로 정하는 것은 이견이 없었으나, 해당 범위에 서비스 제공자, 해당 서비스를 활용(생성 또는 배포·유통 등)하는 사람까지 모두 포함되어야 하는지 여부가 쟁점이 되었다.

이용자 범위를 생성형 인공지능 ‘서비스 이용자’로 정의할 경우, 해당 서비스를 이용하는 사람으로만 범위가 한정될 우려가 깊어 논의되었다. 예를 들어, 생성형 인공지능 서비스 이용자가 만든 콘텐츠를 소비하는 이용자는 해당되지 않고, 범위 밖으로 벗어나게 되는 것이다. 즉, 이용자의 범위를 일반적인 이용자로 할 것인지 생성형 인공지능 서비스 이용자로 좁게 한정할 것인지 명확하게 할 필요가 있었다. 해당 논의는 향후 2차, 3차에까지 지속되었으나 최종 결론으로 본 가이드라인에서 ‘이용자’의 범위는 이 가이드라인의 기초가 되는 방송통신위원회의 「이용자 중심의 지능정보사회를 위한 원칙」(19)에 준하게 정의하기로 하였다. 즉, 인공지능 기반 서비스를 이용하는 이용자(서비스 이용자)를 대상으로 하되, 서비스 이용단계에서 뿐만 아니라 유통단계에서의 이용자 보호 이슈도 함께 고려하는 것이다.

다음으로 생성형 인공지능 서비스 이용자가 어떤 측면에서 권익침해를 받는지, 보호의 필요성이 존재하는지에 대한 쟁점을 논의하였다. 기본적으로 생성형 인공지능 서비스 이용자 보호 이슈는 서비스별로 접근할 필요가 있으며, 개인정보, 저작권 등 기존에 제기된 이슈 이외에 출처정보 고지와 다양성 이슈 등도 고려해야 한다는 데 의견이 모였다. 생성형 인공지능 서비스 이용자에게 대한 권익 침해는 필터버블, 맞춤형화를 가장한 획일화 등 n차 피해가 예상되는 만큼 피해의 유형화 작업이 선행될 필요가 있다고 보았다.

또 다른 의견으로 소비자 관점의 논의가 있었다. 예를 들어, 마케팅을 위해 생성형 인공지능 서비스를 이용하여 맞춤형을 가장한 차별화(불공정 이슈)를 할 경우, 사전 고지가 되었다면 이를 정당한 마케팅을 위한 이용으로 인정할 것인지와 같은 쟁점을 논의한 것이다. 사업자에게 책임이 있다면, API를 받아서 서비스를 제공하는 사업자에게 책임을 부과할

것인지, 파운데이션 모델을 제공하는 사업자에게 책임을 부과할 것인지도 문제가 될 수 있다. 다양성 이슈와 관련하여 기업 자체 편견진단, 피드백 절차가 포함될 필요가 있으며, 보안 차원에서 다양한 공격에 대한 방어적 조치 및 준비가 이루어지고 있는지에 대한 점검도 필요하다는 의견도 주고 받았다. 하나의 서비스 내에서 다양성을 강조하기보다는 거시적 관점에서 다양한 종류의 서비스 제공을 통한 다양성을 보장하는 것도 중요하다는 지적도 있었다.

다른 한편에서는 가이드라인을 구성하는 주요 내용에 규제적 요소가 다수 포함된다면 기업에 과도한 부담을 지울 수 있다는 우려가 논의되었고, 수범자를 스타트업으로 가정할 경우 해당 가이드라인을 준수할 수 있는 수준의 내용인지도 검토해야 한다는 의견을 주고 받았다. 이처럼 연구반 구성원들은 가이드라인의 내용을 논의하는 동안, 생성형 인공지능 서비스 활성화 측면과 이용자 보호 사이의 적절한 균형점을 찾는 노력을 지속하였다.

사업자들에게 사전 질문할 문항을 논의할 때는 우선 생성형 인공지능 서비스 이용자 보호 관련 기존 발표된 가이드라인 및 정책자료 검토를 통해 이슈를 도출하여 정리하였으며, 추가로 사업자의 이용자 정책의 기초와 방향, 서비스 안정성을 위한 노력, 위기사항 시 대응책 여부 등을 주요 문항으로 제하였다.

2. 제2차 가이드라인 연구반 전체회의

생성형 인공지능 서비스 이용자 보호 가이드라인 연구반 제2차 전체회의는 3월 28일(목) 토즈 양재점에서 개최되었다. 지난 제1차 생성형 인공지능 서비스 이용자 보호 가이드라인 연구반 킷오프 이후, 본 전체회의에서는 방송통신위원회 이용자 보호 기본 원칙 및 국내외 관련 법제 동향을 고찰하고, 해당 가이드라인의 최적 모델 및 내용 수준을 논의하고자 하였다. 세부 논의 내용은 다음과 같다.

〈표 3-3〉 제2차 가이드라인 연구반 전체회의 프로그램

시간		내용	비고
15:00~15:10	10'	• 인사말씀	• 천지현 과장(방통위)
15:10~15:30	20'	• 연구반 주요 논의 사항 소개 • 방통위 이용자보호 기본원칙 및 국내 관련 법 제 및 가이드라인 핵심 내용 공유	• 조성은 연구위원(KISDI)
15:30~15:50	20'	• 인공지능 관련 해외 규범 동향 고찰	• 선지원 교수(한양대)
15:50~16:50	60'	• 주요 논의 사항 토론 및 가이드라인 모델·내 용 수준 합의	• 참석자 전원
16:50~17:00	10'	• 향후 일정 공유 및 폐회	• 방통위/KISDI

먼저, 방송통신위원회의 기본원칙들을 살펴보았다. 이 기본원칙들은 방송통신위원회와 정보통신정책연구원이 공동개발한 것으로, 2019년 11월 11일에 발표한 「이용자 중심의 지능정보사회를 위한 원칙」에 이어, 2021년 6월 30일에 발표한 「인공지능 기반 미디어 추천 서비스 이용자 보호 기본원칙」을 다루었다. 이번에 개발하는 생성형 인공지능 서비스 가이드라인은 이 기본원칙들을 기초로 하여 개발할 예정이었기에, 이 원칙들의 목적 및 지향점, 이용자 정의, 기본원칙 등을 공유하고 관련 논의를 하였다.

〈표 3-4〉 방통위의 이용자 보호를 위한 자율 규제

구분	이용자 중심의 지능정보사회를 위한 원칙(2019. 11)	인공지능 기반 미디어 추천 서비스 이용자 보호 기본원칙(2021. 6)
목적 및 지향점	• 신기술의 도입이 초래할 수 있는 기술적, 사회적 위험으로부터 안전한 지능정보서비스 환경을 조성하기 위해, 지능정보사회의 모든 구성원들이 고려할 공동의 기본 원칙을 제시	• 디지털 미디어 플랫폼에서 상용되는 인공지능 기반 추천 서비스의 투명성과 공정성을 제고하기 위하여 추천 서비스 제공자에게 권고되는 자율적인 실천규범으로, 디지털 미디어 콘텐츠의 편향적·차별적 제공을 방지하고 이용자의 권익을 적극적으로 보장함으로써 지능정보사회의 자유민주적 기반을 형성하는 미디어의 가치를 수호 목적

구분	이용자 중심의 지능정보사회를 위한 원칙(2019. 11)	인공지능 기반 미디어 추천 서비스 이용자 보호 기본원칙(2021. 6)
이용자 정의		• 디지털 미디어 플랫폼에서 제공되는 추천 서비스를 이용하거나 이용하려는 자를 말하며 해당 플랫폼을 통하여 미디어 콘텐츠를 소비하는 자 또는 플랫폼 서비스를 이용하여 미디어 콘텐츠를 공급하는 자를 모두 포함
기본원칙	• ① 사람 중심의 서비스 제공 원칙 ② 투명성과 설명가능성 원칙 ③ 책임성 원칙 ④ 안전성 원칙 ⑤ 차별 금지 원칙 ⑥ 참여 원칙 ⑦ 프라이버시와 데이터 거버넌스	• [3대 핵심원칙] ① 투명성 ② 공정성 ③ 책무성 • [5대 실행원칙] ① 이용자를 위한 정보 공개 ② 이용자의 선택권 보장 ③ 자율검증 실행 ④ 불만 처리 및 분쟁 해결 ⑤ 내부 규칙 제정
기타	• [이용자 보호를 위한 공동의 노력] - 지능정보사회의 구성원들은 공동의 기본 원칙에 입각하여 지능정보사회의 기본 가치를 수호할 수 있도록 자율적인 노력을 지속한다. -지능정보사회 이용자 보호를 위한 실질적이고 지속적인 논의를 위해 지능정보사회의 구성원들이 참여하는 협의회를 구성하여 운영한다.	• [기본원칙 이행을 위한 정부의 역할] ① 서비스 제공자 지원 ② 이용자 역량 강화

자료: 연구진 작성

다음으로 생성형 인공지능 서비스의 이용자에 대한 개념과 범위를 설정하기 위해 국내 인공지능 관련 법제 및 가이드라인 핵심 내용을 탐구하면서 다양한 논의가 이루어졌다. 먼저 범정부 인공지능 윤리기준을 살펴보았다. 이 윤리기준은 과학기술정보통신부와 정보통신정책연구원이 공동 개발하여 2020년 12월 범정부 인공지능 윤리기준으로 발표한 것이다. 주요 내용은 <표 3-5>에서 살펴볼 수 있다.

〈표 3-5〉 범정부 인공지능 윤리기준

구분	사람이 중심이 되는 인공지능 윤리 기준(2020. 12)
배경 및 필요성	• 새로운 지능정보기술의 장점을 극대화하면서도 새로운 기술의 특성에 따라 나타날 수 있는 예기치 못한 기술적, 사회적 위험을 최소화할 수 있도록 하는 고려가 필요
목적 및 지향점	• 인공지능 시대 바람직한 인공지능 개발·활용 방향을 제시하기 위함 • ① 모든 사회 구성원이 ② 모든 분야에서 ③ 자율적으로 준수하며 ④ 지속 발전하는 윤리기준을 지향 (① 인공지능 개발에서 활용에 이르는 전 단계에서 정부·공공기관, 기업, 이용자 등 모든 사회 구성원이 참조하는 기준, ② 특정 분야에 제한되지 않는 범용성을 가진 일반원칙으로, 이후 각 영역별 세부 규범이 유연하게 발전해나갈 수 있는 기반 조성, ③ 구속력 있는 ‘법’이나 ‘지침’이 아닌 도덕적 규범이자 자율규범으로, 기업 자율성을 존중하고 인공지능 기술발전을 장려하며 기술과 사회변화에 유연하게 대처할 수 있는 윤리 담론을 형성, ④ 사회경제, 기술 변화에 따라 새롭게 제기되는 인공지능 윤리 이슈를 논의하고 구체적으로 발전시킬 수 있는 플랫폼으로 기능)
기본 원칙	• (3대 기본원칙) ‘인간성(Humanity)’을 구현하기 위해 인공지능의 개발 및 활용 과정에서 ① 인간의 존엄성 원칙, ② 사회의 공공선 원칙, ③ 기술의 합목적성 원칙을 준수해야 함 • (10대 핵심요건) 3대 기본원칙을 실천하고 이행할 수 있도록 인공지능 개발 및 활용의 전 과정에서 ① 인권 보장, ② 프라이버시 보호, ③ 다양성 존중, ④ 침해금지, ⑤ 공공성, ⑥ 연대성, ⑦ 데이터 관리, ⑧ 책임성, ⑨ 안전성, ⑩ 투명성의 요건이 충족되어야 함

자료: 연구진 작성

이어 국내 사업자들의 인공지능 윤리기준도 살펴보았다. 사업자들은 이용자의 신뢰와 서비스 품질 관리를 위해 자체 윤리기준을 마련하고 이를 대외 공표하여 사업자의 노력을 널리 알리고자 하였다.

〈표 3-6〉 국내 기업의 인공지능 윤리기준

구분	목적 및 지향점	기본 원칙
카카오 알고리즘 윤리헌장 (2018)	• 카카오 사내에서 인공지능 관련 연구 시 지향되어야 할 윤리 원칙들을 제시 • 인공지능 기술 연구자가 고려해야 할 7개의 윤리원칙들을 제시	• 카카오 알고리즘의 기본 원칙 • 차별에 대한 경계 • 학습 데이터 운영 • 알고리즘의 독립성 • 알고리즘에 대한 설명 • 기술의 포용성 • 아동과 청소년에 대한 보호
네이버 인공지능 윤리준칙 (2021)	• 네이버의 모든 구성원이 인공지능 개발과 이용에 있어 준수해야 할 원칙 • 윤리준칙 공개와 함께 실천 사례 축적 및 개선을 위한 학계와의 지속적 협력을 통해 준칙을 더욱 구체화하고 개선해나갈 것을 밝힘	• 사람을 위한 AI 개발 • 다양성의 존중 • 합리적인 설명과 편리성의 조화 • 안전을 고려한 서비스 설계 • 프라이버시 보호와 정보 보안
SK텔레콤 AI 추구 가치 (2021)	• 사람과 인공지능이 공존하는 행복한 세상을 만드는 것, 이를 위해 7개의 인공지능 추구 가치를 선포	• 사회가치 • 무해성 • 기술안정성 • 공정성 • 투명성 • 사생활보호 • 지속혁신
LG AI 윤리 원칙 (2022)	• 신뢰할 수 있는 인공지능을 개발하기 위한 필수적인 가치를 비롯해 LG의 경영 이념인 ‘고객을 위한 가치창조’와 ‘인간 존중의 경영’을 기반으로 선정	• 인간존중(Humanity) • 공정성(Fairness) • 안전성(Safety) • 책임성(Accountability) • 투명성(Transparency)

〈표 3-7〉은 기개발된 공공기관의 생성형 인공지능 관련 활용 가이드라인을 정리한 것이다. 앞서 언급하였듯이, 기개발된 생성형 인공지능 가이드라인들은 활용단계에서 이용자가 지켜야 할 지침 중심으로 개발되었다.

〈표 3-7〉 기존 생성형 인공지능 가이드라인에서의 이용자 개념과 주요 내용

	특징
서울시 생성형 인공지능 윤리가이드 (23. 11)	• 이용자가 생성형 인공지능을 활용할 때 필요한 사항을 설명하는 가이드라인 • 이용자의 질문으로부터 콘텐츠가 생성되기 때문에, 이용자가 생성형 인공지능 활용 목적을 결정하는 것으로 정의, 또한 이용자 의도에 따라 생성형 인공지능의 활용 범위가 넓어질 수 있다고 봄 • 이용자 권리와 이익을 보호하고 활용하는 차원에서 인공지능을 개발하고 확대하기 위한 이용자 중심의 윤리 가이드라인이 필요하다는 점을 도출하였으며, 기본원칙으로 ‘이용자 중심’을 제시함 • 이용자가 갖추어야 할 역량, 책임성, 데이터 권리와 책임, 저작권 고려, 학습 데이터 안전성, 인공지능 윤리 소양, 공공성, 지속가능성, 보안성 등 9대 핵심 요건 제시
국가정보원 챗GPT 등 생성형인공지능 활용 보안 가이드라인 (23. 6)	• 안전한 생성형 인공지능 기술 사용 가이드라인으로서, 각급 기관의 정보화·정보보안 담당자들이 보다 효율적이고 안전하게 서비스를 이용하도록 하는 데 그 목적이 있음 • FAQ 형식으로 구성하였으며, ① 서비스 사용 주의사항(서비스 접근, 계정 관리법), ② 서비스와의 대화 시 주의사항, ③ 인공지능 모델 플러그인 사용 주의사항(올바른 서비스 플러그인 사용·관리, 개인정보·민감정보 처리), ④ 인공지능 모델 확장 프로그램 사용 주의사항 ⑤ 인공지능 모델 생성기반 공격 대처 방안 등을 포함
행정안전부 챗GPT 활용방법 및 주의사항안내 (23. 5)	• 행정기관 직원들이 생성형 인공지능 서비스 사용 시 주의할 점 등 안내 • 챗GPT 개요, 챗GPT 활용가능 분야 및 활용예시, 챗GPT 문제점(한계)과 활용 시 주의사항
방통위-NIA 생성형인공지능 윤리 가이드북 (23. 12)	• 생성형 인공지능을 윤리적이고 생산적으로 활용하는 데 참고 • 저작권, 책임성, 허위조작정보, 개인정보·인격권, 오남용에 관한 내용을 Q&A 형식으로 안내하고 있음
한국인터넷 자율정책기구 (KISO) 챗봇 윤리 가이드라인	• 챗봇 개발자, 운영자, 이용자가 준수해야 할 기본원칙 제시(주체별로도 제시) • (1) 인간의 존엄성 및 권리 존중 원칙(이용자의 인격과 존엄성을 존중, 관련 법령 준수 및 이용자 권리와 윤리 가치 존중) (2) 프라이버시 보호 및 정보보안 원칙(프라이버시 및 개인정보를 보호, 개인정보 오남용 및 무단 공유 금지) (3) 다양성 존중 원칙(동등한 접근성, 불공정한 편향 금지, 이용자 차별 금지, 아동·청소년 및 사회 약자 보호 및 존중) (4) 투명성 원칙(대화를 개시하기 전에 이용자에게 챗봇임을 알릴 것, 필요한 경우 챗봇의 용도와 특성을 이용자가 알기 쉽게 설명) (5) 책임 원칙(챗봇 서비스의 지속가능성과 사회적 영향을 고려할 것, 이용과정 중 발생한 문제에 대한 책임 및 책무성 체계를 갖추도록 권고, 챗봇과 이용자 상호 간의 지속가능한 상생 노력 제안)

자료: 연구진 작성

추가로 공공기관의 영역별 인공지능 서비스 관련 가이드라인도 살펴보았다. 생성형 인공지능 서비스는 아니지만, 자료 정리 및 논의를 하면서 생성형 인공지능의 특수성에 따른 이슈뿐만 아니라 모든 인공지능 기술 및 서비스에 해당하는 이슈들이 생성형 인공지능 서비스에도 여전히 관련되는 만큼, 기존의 인공지능 기술/서비스에 관련한 자율규범을 폭넓게 살펴볼 필요가 있었다. 다음은 연구반에서 공유되고 논의되었던 자료들이다. 국토교통부 자율주행자동차 윤리 가이드라인(〈표 3-8〉), NIA 지능정보사회 윤리 가이드라인과 윤리헌장(〈표 3-9〉), 교육부 교육분야 인공지능 윤리 원칙(〈표 3-10〉), 국가인권위원회 인공지능 개발과 활용에 관한 인권 가이드라인(〈표 3-11〉)을 참고하면 된다.

〈표 3-8〉 국토교통부 자율주행자동차 윤리 가이드라인

구분	국토교통부 자율주행자동차 윤리 가이드라인(2020)
목적 및 지향점	• 본 가이드라인은 정해진 조건에서 운전에 전혀 개입하지 않거나, 사람이 직접 운전에 관여하지 않는 레벨4 이상 완전자율주행시스템이 장착된 자동차를 설계·제작·관리 운행하려는 자와 사용자(소비자) 등을 대상으로 함 • 자율주행자동차의 도입으로 편리하고 자유로운 이동권 보장 및 교통사고 피해 감소 등의 긍정적 효과를 증진시키고, 자율주행자동차의 도입으로 발생할 수 있는 부정적 효과를 감소시키는 동시에 책임 소재를 가릴 수 있는 방향을 제시하며, 관련 행위주체들이 자율주행자동차의 기본 가치에 따라 각각의 행동원칙을 준수할 것을 권고하기 위함
기본원칙	1. 자율주행자동차는 인간의 존엄성, 국제법적으로 인정된 인권과 자유, 프라이버시 및 문화적 다양성을 존중하고, 인간을 성별, 나이, 인종, 장애 등을 이유로 차별하지 않으며, 인간의 법과 관습에 의한 판단과 통제에 따르도록 설계, 제작, 관리되어야 한다. 2. 자율주행자동차는 인간의 행복과 이익의 증진을 위한 수단으로서 인간의 안전하고 편리하며 자유로운 이동권을 보장하고, 타인의 권리와 자유를 침해하지 않도록 설계, 제작, 관리되어야 한다. 3. 자율주행자동차는 자동차 사고로 인해 발생할 손실을 최소화하고, 무엇보다 인간의 생명을 우선하도록 설계, 제작, 관리되어야 한다. 또한, 손실을 최소화하는 과정에서 인간을 성별, 나이, 종교 등 개인적 차이 등을 이유로 차별하지 않고, 교통 약자를 고려하는 방식으로 작동하도록 설계, 제작, 관리되어야 한다. 4. 자율주행자동차는 지구 온난화, 교통체증 등을 고려하여 지속 가능한 발전이 이루어지도록 설계, 제작, 관리, 운영되어야 한다.

자료: 연구진 작성

〈표 3-9〉 NIA 지능정보사회 윤리 가이드라인과 윤리헌장

구분	NIA 지능정보사회 윤리 가이드라인과 윤리헌장(2018)
목적 및 지향점	• 인간 중심의 지능정보사회를 구현하기 위해, 기능정보기술 및 서비스 개발자·공급자의 책임윤리 강화 및 이용자의 오남용 방지를 위한 지침 제공
기본원칙	1. 공공성 - 많은 이에게 도움을 주고 모든 인류의 혜택을 위해 공유되어야 한다. 2. 책무성 - 사고 등의 책임 분배를 명확히 하고, 안전 관련 정보 공유, 이용자 권익보호 등 사회 의무 수행에 최선을 다해야 한다. 3. 통제성 - 인간의 제어 가능성 및 오작동에 대한 대비책 마련과 이용자 이용 선택권을 최대한 보장해야 한다. 4. 투명성 - 기술 개발, 서비스설계, 제품기획 등 의사결정 과정에서 이용자·소비자·시민 등의 의견을 반영하도록 노력해야 하며, 이용 단계에서 예상되는 위험과 관련한 정보를 공개·공유하고, 개인정보 처리의 전 과정은 적절하게 이루어져야 한다.

자료: 연구진 작성

〈표 3-10〉 교육부 교육분야 인공지능 윤리 원칙

구분	교육부 교육분야 인공지능 윤리 원칙(2022)
목적 및 지향점	• 인공지능의 교육적 활용 증대 등 변화에 따라 교육 현장에 도입되는 인공지능의 역기능, 부작용을 최소화하기 위한 선제적 자율규제 마련 필요 • 인공지능의 이용자인 교육당사자(수요자)*가 개발자(공급자) 등에게 요구할 수 있는 권리로서의 윤리규범 * 교수자, 학습자, 학부모 • 사람의 전 생애에 걸쳐 전인적 성장을 최고 가치로 삼으며, 사람의 인격을 존중하고 개성을 중시하여 사람의 능력이 효과적으로 발휘될 수 있도록 제 공되어야 함
기본원칙	[10대 원칙] 1. 인간성장의 잠재성을 이끌어낸다. 2. 학습자의 주도성과 다양성을 보장한다. 3. 교수자의 전문성을 존중한다. 4. 교육당사자 간의 관계를 공고히 유지한다. 5. 교육의 기회균등과 공정성을 보장한다. 6. 교육공동체의 연대와 협력을 강화한다. 7. 사회 공공성 증진에 기여한다. 8. 교육당사자의 안전을 보장한다. 9. 데이터 처리의 투명성을 보장하고 설명 가능해야 한다. 10. 데이터를 합목적적으로 활용하고 프라이버시를 보호한다.

자료: 연구진 작성

〈표 3-11〉 국가인권위원회 인공지능 개발과 활용에 관한 인권 가이드라인

구분	국가인권위원회 인공지능 개발과 활용에 관한 인권 가이드라인(2022)
목적 및 지향점	인공지능을 개발할 때부터 개인의 삶과 사회적 공익에 기여할 수 있도록 개발하고, 인간의 존엄성과 자기결정권의 보장 및 차별받지 않을 권리 등 인권에 기반하여 활용하며, 인공지능 기술로 발생할 수 있는 인권침해의 예방 및 권리구제 절차를 마련하는 것이 중요
기본원칙	1. 인간의 존엄성 - 인공지능은 인간의 존엄과 가치 및 행복을 추구할 권리에 부합하는 방향으로 개발 및 활용되어야 하며, 개인의 선택과 판단 및 결정을 강요하거나 자율성을 침해할 수 없다. 2. 투명성과 설명 의무 - 인공지능 기술을 활용한 판단 과정과 그 결과에 대해 적절하고 합리적인 설명이 보장되어야 한다. 3. 자기결정권의 보장 - 정보주체는 인공지능 서비스가 언제, 어디서 자신의 개인정보를 수집하고, 어떻게 개인정보를 처리하여 사용, 보관, 삭제하는지에 대해 알고 참여할 권리가 있다. 인공지능을 활용한 의사결정의 내용과 관련성이 없거나, 부정확한 데이터에 기반한 의사결정이 이루어지지 않도록 데이터의 정확성, 완전성, 최신성을 보장해야 한다. 4. 차별 금지 - 다양성과 대표성 반영, 데이터 관리 5. 인공지능 인권영향평가 시행 - 국가는 인공지능의 개발과 활용에 있어서 인권침해와 차별의 가능성 및 정도, 영향을 받은 당사자의 수, 사용된 데이터의 양 등을 고려하여 인권영향평가를 실시해야 한다. 6. 위험도 등급 및 관련 법·제도 마련 - 국가는 공공기관과 민간이 개발하고 활용하는 인공지능이 인권과 책임성을 보장할 수 있도록 관련 법률과 감독 체계를 보완해야 하며, 구체적인 지침과 정책을 마련해야 한다.

자료: 연구진 작성

3. 제3차 가이드라인 연구반 전체회의

제3차 전체회의는 4월 19일(금) 상연재 서울역점에서 있었다. 3차 회의 전에 2차 회의 때까지 논의되었던 내용을 바탕으로 연구반 위원들에게 생성형 인공지능 서비스 이용자

보호 이슈에 대한 쟁점 및 문제해결을 위한 원칙·기준·구체적 실천방안에 대한 서면 의견을 먼저 받았다.

서면의견서는 먼저 ‘생성형 인공지능 서비스 이용자 보호 이슈에 대한 쟁점’을 적고, 해당 쟁점을 다룰 때 필요한 ‘문제해결을 위한 원칙·기준’을 제시하며, 이어 ‘구체적 실천방안’에 관한 의견을 작성하도록 구성되었다. ‘문제해결을 위한 원칙·기준’은 앞서 다루었던 방송통신위원회의 기본원칙을 기조로 하되, 그 밖의 문헌들도 참조하도록 하였다.

실무진에서 수집한 서면의견서들을 ‘문제해결을 위한 원칙·기준’별로 유형화하여 재구성하였다. 그 결과 타인의 권리 및 사생활 침해, 투명성 및 설명가능성, 저작권 침해, 이용자 선택권 보장, 책임소재 불분명, 불법 콘텐츠 유통, 인간의 존엄성 훼손, 다양성 존중, 고지의무 총 9개의 주제로 분류되었으며, <표 3-12>에서 <표 3-20>까지가 주제별 주요 내용을 요약한 것이다.

<표 3-12> 주제1. 타인의 권리 및 사생활 침해(악의적 활용)

문제·이슈	• 사생활 침해나 명예훼손 등 타인의 권리를 침해하는 정보 또는 불법 정보가 출력 데이터(output data)로 생성되는 경우 • (타인권리 침해 정보 또는 불법정보에 해당하지는 않으나) 비윤리적·편향적 정보가 출력 데이터(output data)로 생성되는 경우 • 생성형 인공지능을 활용하는 과정에서 타인의 명예와 프라이버시를 침해하거나 상업적으로 이용할 경우의 문제
문제 해결을 위한 원칙·기준	• (타인권리 침해 정보 또는 불법정보 관련) 책임성 원칙, 프라이버시와 데이터 거버넌스(이용자 중심의 지능정보사회를 위한 원칙) ※ 지능정보화 기본법 제62조(지능정보사회윤리) 중 “책임성” 참조 • (비윤리적·편향적 정보 관련) 차별 금지 원칙, 사람 중심의 서비스 제공 원칙(이용자 중심의 지능정보사회를 위한 원칙) ※ 지능정보화 기본법 제62조(지능정보사회윤리) 중 “인간의 존엄과 가치 존중” 참조
구체적 실천방안	• 이용자가 타인의 권리 침해 정보/불법 정보/비윤리적·편향적 정보를 프롬프트에 입력하는 경우 경고문구 제시, 일정 시간 차단 (KISO 챗봇 윤리 가이드라인 중 “4. 개발자 준수사항”, “5. 운영자 준수사항” 참조) • 필터링 및 순화 알고리즘 적용(인공지능 윤리 국가표준) • 이용자 신고 프로세스 및 의견 수렴 창구 구현 • 인공지능 시스템 책임자 지정 및 주기적 모니터링 • 기업의 인공지능 윤리 거버넌스 마련

실제 사례 혹은 현황	[실제 사례] • 주요 기업(OpenAI, 네이버 등)은 이용약관·정책 등을 마련하여 타인의 권리 침해 정보/불법정보/비윤리적·편향적 정보 생성 등 악의적으로 인공지능 서비스를 사용하는 것을 금지하고 있으며, 신고 프로세스 및 의견 수렴 창구를 구축하고 있음 • 상업적 목적을 위해 유명인이나 유명 콘텐츠를 비슷하게 만들어 이용하는 업체들 & 유명 디자인 분위기를 카피한 작품을 과제로 제출 [참조 법률] • 정보통신망 이용촉진 및 정보보호 등에 관한 법률 제44조(정보통신망에서의 권리보호) 제2항, 제44조의2(정보의 삭제 요청 등) 제2항, 제44조의7(불법 정보의 유통금지 등) 제1항, 제44조의8(대화형정보통신서비스에서의 아동 보호), 제44조의9(불법촬영물 등 유통방지 책임자) 제1항 • 전기통신사업법 제22조의5(부가통신사업자의 불법촬영물 등 유통방지) 제1항 및 제2항, 제32조(이용자 보호) 제1항 • 방송통신위원회의 설치 및 운영에 관한 법률 제21조(심의위원회의 직무) 제4호
----------------	---

자료: 연구진 작성

〈표 3-13〉 주제2. 투명성 및 설명가능성

문제·이슈	• 생성형 인공지능 모델 결과물에서 불공정·편견·차별적 내용 도출 이슈 • 생성형 인공지능은 결과 도출의 과정이 블랙박스이기 때문에 사용자들은 어떤 과정을 거쳐 그러한 결과를 도출하였는지 알 수 있는 방법이 없음 • 이용자를 위한 정보 생성 결과물을 소비자가 절대적인 것으로 오인하고 활용할 때 생기는 문제(이용자가 '생성형 인공지능'의 의미와 그 알고리즘을 정확하게 파악하지 못해 생기는 문제)의 가능성
문제 해결을 위한 원칙·기준	• 투명성 및 설명가능성 ① 언제: 항상/사용자 요구 시에만 ② 어떻게: 대화창 내/대화창 밖 ③ 무엇을: 이유(결과 도출의 이유)만, 아닌 이유(다른 결과를 도출하지 않은 이유), what if(다른 상황을 가정하여 또 하나의 다른 답을 제시)
구체적 실천방안	• 학습데이터 종류 및 범위, 생성 콘텐츠의 출처 표시(*파운데이션 모델을 가져와 사용하는 서비스 제공자에게 해당 안됨) • 불공정·편견·차별을 방지하기 위한 자체 조치 사항 정보 • 이용자가 서비스를 이용하는 과정에서 발생하는 불공정·편견·차별 내용에 대한 책임소재 • 서비스 제공자는 사용하는 파운데이션 모델에 대한 정보 및 개발사가 밝힌 모델의 학습데이터, 학습주기, 아용자와의 상호작용과정에서 발생한 정보를 학습데이터로 사용하는지 등의 투명성 정보를 고지

구체적 실천방안	• 생성형 인공지능의 작동원리와 한계, 그리고 생성형 인공지능의 결과물을 이용하는 과정에서 이용자가 저야 할 책임까지 고지
실제 사례 혹은 현황	• EU AI법에 범용인공지능 모델 학습데이터의 주요 내용 공지, 훈련·평가 프로세스와 결과 등에 관한 기술 서류를 인공지능시스템제공자에게 공개하는 등의 투명성 조치 권고 • 수업 중 학생들이 생성형 인공지능의 결과물을 정답으로 생각하고 이의제기하는 경우가 있음

자료: 연구진 작성

〈표 3-14〉 주제3. 저작권 침해

문제·이슈	• 이용자가 사용하는 콘텐츠가 인간의 창작물인지 인공지능이 만든 콘텐츠인지 구분하지 못하는 이슈 • 저작권 침해 데이터로 학습한 생성형 인공지능을 이용하여 콘텐츠를 생성하였는데, 그 콘텐츠에 저작권 시비 문제가 발생하는 경우 * 애초 파운데이션 모델의 학습데이터 관리가 필요한 문제 • 이용자의 명령글이 저작권 침해를 유도하는 경우 (예: 사진작가A의 ***사진을 ...해줘.)
문제 해결을 위한 원칙·기준	• 안전성, 이용자를 위한 정보 공개, 이용자의 선택권 보장 • 침해금지, 저작권
구체적 실천방안	• 생성된 콘텐츠의 창작자 구분 표시(예: 인공지능 워터마크) • 개발사 책임 문제이므로 개발사 책임 범위를 제시 • 저작권 침해를 유도하는 명령어를 작성하지 못하게 필터링할 기술적 방법 • 저작권 침해 가능성이 있는 사안에 대한 이용자 고지
실제 사례 혹은 현황	• 페이스북·인스타그램은 2024년 5월부터 인공지능 콘텐츠에 라벨을 표기할 예정 • EU AI법에서 지적재산권법 준수를 위한 정책 수용을 권고

자료: 연구진 작성

〈표 3-15〉 주제4. 이용자 선택권 보장

문제·이슈	이용 과정에서 생성한 콘텐츠를 이용자가 명백히 인지하지 못한 채 인공지능 학습데이터로 활용되는 이슈
문제 해결을 위한 원칙·기준	- 이용자의 선택권 보장 - 프라이버시·개인정보 침해 금지
구체적 실천방안	- 생성형 인공지능 이용 과정에서 생성된 콘텐츠를 인공지능 학습데이터로 활용할지 말지를 이용자가 선택하게 하는 고지 및 절차 마련 * 일부 기업은 사용하지 않는다고 고지하거나 거부가능하게 함(OpenAI, 구글)
실제 사례 혹은 현황	- 주요 기업(OpenAI, 구글 등)은 ‘프라이버시정책(Privacy policy)’ 혹은 ‘개인정보 처리방침’(일부) 정보를 수집, 보관, 학습에 활용할 수 있음을 고지하고 있음 * 현재 이용 과정에서 생성된 콘텐츠를 수집·보관·학습에 활용을 거부할 수 있는 선택권 절차는 없는 것으로 보임

자료: 연구진 작성

〈표 3-16〉 주제5. 책임소재 불분명

문제·이슈	- 생성형 인공지능 플랫폼이나 그 결과물을 이용하는 과정에서 문제(경제적, 신체의적 피해 등)가 생겼을 때, 어디에서 문제를 해결해야 하는지를 모를 경우 발생 가능 - 서비스 이용 과정에서 권리를 침해당했다고 판단되었을 때, 피해를 호소할 방법의 필요
문제 해결을 위한 원칙·기준	- 불만처리 및 분쟁 해결 - 침해 금지
구체적 실천방안	- 어떤 상황이든 이용자들의 문의 및 불만처리, 분쟁해결을 위한 접근가능한 채널을 공개할 것 (기본적으로 상호 점점에 있는 단위끼리 해결 필요) - 이용자와의 1차 상호 점점에 있는 기관에서 1차로 책임을 지도록 하는 원칙이 필요 - 권리침해를 당했다고 판단되었을 때 구제 절차를 요청할 수 있는 방법을 제공할 필요
실제 사례 혹은 현황	- 관여하는 당사자가 여럿인 다른 서비스(예: 오픈마켓)에서도 서로 책임을 회피하여 소비자가 피해를 해결할 수 없는 경우가 있음 (예: 쇼핑 플랫폼-사업자, 쇼핑물 입점 브랜드, 판매자, 배달업자 등 책임을 미루는 경우) - 고객문의 같은 기존 시스템을 이용하게 하는 경우(위튼), 따로 신고절차를 제공하는 경우(구글, 네이버, MS), 혹은 아예 없는 경우

자료: 연구진 작성

〈표 3-17〉 주제6. 불법 콘텐츠 유통

문제·이슈	• 이용자가 불법 콘텐츠*를 생성하여 유포하는 경우 * 명예훼손, 음란물, 보이스 및 화면 피싱, 악성코드 생성 및 유포 등
문제 해결을 위한 원칙·기준	• 인권 존중, 침해 금지
구체적 실천방안	• 이용자 간 신고 절차 또는 상시 모니터링 • 기술적 조치로 특정 단어, 주제 관련 입력 금지 조치 및 고지 (예: 얼굴 이미지 사진 업로드 거부)

자료: 연구진 작성

〈표 3-18〉 주제7. 인간의 존엄성 훼손

문제·이슈	• 막말로봇으로 유명한 ‘소피아’는 인간을 멸망시킬 것이다, 인간의 아이를 임신하고 싶다 등 인간과 대치될 수 있는 발언들을 한 바 있음
문제 해결을 위한 원칙·기준	• 인간의 존엄성 및 권리 존중
구체적 실천방안	• 개발의 단계에서 인간의 존엄성이 침해될 수 있는 표현이나 내용을 담지 않도록 플랫폼들이 자율적이고 상시적으로 체크함

자료: 연구진 작성

〈표 3-19〉 주제8. 다양성 존중

문제·이슈	• 생성형 인공지능 모델이 지나치게 개인의 선호를 정확하게 파악하여 결과물을 제공할 경우, 필터버블의 문제가 발생할 수 있음
문제 해결을 위한 원칙·기준	• 다양성 ① 콘텐츠 다양성: 기본적으로 생성하는 콘텐츠에 있어서, 다양성을 증진시키도록 최대한 다양한 내용들을 제공한다(양적 다양성, 종류의 다양성, 비차별성 등). ② 선택 다양성: 사용자들이 다양한 선택지를 가질 수 있도록 편향되지 않은 콘텐츠를 생성한다.
구체적 실천방안	• 플랫폼 사업자가 자율적으로 다양성에 대해 체크 • 다양성 지표를 개발하여 사용자에게 공지

자료: 연구진 작성

〈표 3-20〉 주제9. 고지의무

문제·이슈	• 부정확하거나 사실이 아닌 정보를 마치 사실인 양 제시하는 문제
문제 해결을 위한 원칙·기준	• 책임성, 데이터관리, 정확성, 안전성
구체적 실천방안	• 생성 콘텐츠의 원문정보(링크) 제공을 권고
실제 사례 혹은 현황	• 현재 기업들 조치에서는, 생성형 인공지능의 결과물이 부정확할 수 있다는 점을 서비스 화면에 고지함 • 또한, 생성 콘텐츠의 완전성, 무결성, 적절성 등을 보증하지 않는다는 면책 조항을 두거나 약관 등에 고지함(OpenAI, 구글, 네이버, MS 등) • EU AI법에서 딥페이크 생성 인공지능시스템에 대해 '인공지능이 생성한 콘텐츠'임을 밝히거나 이용자가 인공지능(기계)과 상호작용하고 있다는 것을 고지 및 공개해야 하는 투명성 의무 부과

자료: 연구진 작성

앞서 확인한 대로 연구반 위원들의 의견들을 중복 의견을 제외하고 주제별로 분류한 자료를 바탕으로 3차 회의를 진행하였다. 주요 논의된 것들을 정리하면 다음과 같다.

먼저 생성형 인공지능의 활용이 증가함에 따라 명예, 프라이버시, 저작권 등 타인의 권리를 침해할 가능성이 커지고 있다는 의견이 나왔다. 관련하여 이를 상업적으로 이용하는 경우 법적 문제가 발생할 수 있어 이를 방지하기 위한 최소한의 지침이 필요하다는 데 대체로 동의하였다. 또, 생성형 인공지능과 관련된 권리로 퍼블리시티권이 최근 부정경쟁방지법에 포함되었다는 정보와, 아직 법령이 되지 않은 상태에서 타인을 식별할 수 있는 기준에 대한 논의가 한동안 지속될 것이라는 현황을 공유하였다. 이처럼 이슈가 있고 해결방안이 논의되고 있더라도 명확하나 법령이나 사회적 합의가 마무리된 기준이 되지 않은 상태에 있다면 사업자에게 책무를 강권하는 데 한계가 있다는 점도 다루어졌다. 이번 가이드라인의 수위를 어느 정도로 하는 것이 가장 합리적인지에 대한 고민이었다.

또 다른 이슈로, 서비스 활용 단계에서 이용자 본인에 대한 권리 이슈뿐만 아니라 이용자가 해당 콘텐츠를 활용함으로써 발생가능한 타인의 권리를 침해가능성도 포괄한 지침이 필요한지에 대해 집중 논의되었다. 이는 1차 회의 때 이용자 범위를 어떻게 규정할 것인가와 이어지는 주제였다. 이때 참조해볼 수 있는 조항으로 「정보통신망법」 제44조의8⁴⁾이 제

시되었다. 이용자 보호를 위한 기존 사례를 참고하여 생성형 인공지능의 특성을 반영한 보호 방안을 설계해야 한다는 기준에 서로 동의하고, 논의를 이어갔다.

이용자가 입력하는 데이터의 수집 및 활용과 관련해서는 요금 체계 등 다양한 옵션을 고지하여 이용자가 가입 시 서비스 제공자 측에서 요구하는 정보를 명확히 알 수 있도록 해야 한다는 의견이 제시되었다. 이때 동의 여부를 선택할 수 있는 시스템을 도입하되, 선택권이 제한되지 않도록 주의해야 한다는 점도 지적되었으며, 또한 서비스 이용 과정에서 문제가 발생할 경우, 이를 해결할 수 있는 상담 창구나 신고 절차 등을 명확히 제공하여 이용자의 접근성을 보장해야 한다는데도 합의하였다. 그동안 법령 제정때마다 지적되었던 국내외 사업자 규제 불균형 문제를 고려하여 국내외 사업자가 동일한 수준의 대응체계를 마련할 수 있도록 해야 한다는 것도 언급되었다. 투명성 의무에 대해서는 좀 더 구체적으로 논의되었는데, 예를 들어 생성형 인공지능 서비스의 신뢰성을 제고하기 위해 데이터 사용 방식, 파운데이션 모델의 연결, 파인 튜닝 과정 등을 투명하게 공개해야 하며, 결과물에 대한 오류 가능성을 지속적으로 점검하여 잘못된 정보 생성 방지를 위한 노력을 강화해야 한다는 내용을 담자는 것이었다.

저작권과 개인정보 이슈는 소관 부처가 따로 있는 만큼 해당 부처의 법령이나 가이드라인을 따르도록 안내하는 것이 바람직하다는 데 모두 동의하였다. 따라서 해당 문제들은 생성형 인공지능 서비스 관련 이슈로 간단히 다루면서 이 가이드라인에서는 관련 부처의 내용을 담아내는 형식으로 합의하였다.

4. 제4차 가이드라인 연구반 전체회의

4차 전체회의는 6월 18일(화) 상연재 서울역점에서 개최되었다. 그동안 실무반에서는 초안 작성에 집중하였다. 초안 작성을 위해 5월 회의를 생략하고 6월에 4차 회의를 진행하게 된 것이다. 이 회의에서 위원들은 가이드라인 초안에 대해 검토한 의견을 나누었으며, 위원들은 대략 용어의 모호성, 낮은 실행가능성, 한계점 등을 지적하였다. 또 이를 보완하기

4) 「정보통신망법」〈제44조의8(대화형정보통신서비스에서의 아동 보호)〉

정보통신서비스 제공자는 만 14세 미만의 아동에게 문자·음성을 이용하여 사람과 대화하는 방식으로 정보를 처리하는 시스템을 기반으로 하는 정보통신서비스를 제공하는 경우에는 그 아동에게 부적절한 내용의 정보가 제공되지 아니하도록 노력하여야 한다.

위한 논의가 계속되었는데 모호한 용어를 명확히 하는 표현이나 실행가능성을 높이는 기술적 조치 등을 집중적으로 논의하였다.

먼저 항목 분류의 모호성에 대한 논의가 진행되었다. 초안의 항목들이 명확히 구분되지 않았고, 일부 내용이 중복되어 분류의 애매함을 초래하고 있다는 지적이었다. 다음으로, 청소년 유해 콘텐츠와 불법 콘텐츠에 대한 정의와 구분이 명확하지 않아서 인공지능 생성한 콘텐츠에 대한 책임 소재와 권리 귀속이 불분명한 상황이라는 의견도 있었다. 이어 제안된 실행 방안 중 일부는 현실적으로 실행가능성이 낮으며, 인공지능의 할루시네이션 문제와 생성 콘텐츠의 신뢰성 저하를 근본적으로 극복하는 것은 기술적 한계 때문에 어렵다는 것도 지적되었다. 관련 의견은 <표 3-21>에 정리하였다.

<표 3-21> 가이드라인 초안에서 논의된 문제점

문제	구체적 논의 내용
각 항목 분류의 모호성	• 시행지침의 항목들이 명확히 구분되지 않고, 일부 내용이 다른 항목과 유사함(예: '1번-인격권 침해'와 '5번-콘텐츠 활용 시 문제', '6번-유통 문제'에서 인격권 침해와 프라이버시 침해 등의 내용이 중복됨) • 생성 과정에서 발생하는 문제와 생성된 콘텐츠 활용 시 발생하는 문제가 혼재되어 있음 • 생성형 인공지능 산출물의 범위를 명확히 정의해야 함 • 교육 등 일상적인 사용과 상업적 사용을 구분할 필요가 있음 • 인공지능이 편향된 데이터를 학습하는 것과 사용자의 선택적 노출을 통해 형성되는 필터 버블을 구분해야 함
청소년 유해 콘텐츠와 불법 콘텐츠의 경계 모호	• 청소년 유해 콘텐츠와 불법 콘텐츠에 대한 정의와 구분이 명확하지 않음 • 불법 정보가 아닌 청소년 유해 정보는 다른 방식으로 규제해야 함 • 인공지능이 생성한 콘텐츠에 대해 등급을 매기는 접근 방식이 가능할지에 대해 의문이 제기됨
책임 및 권리 범위의 모호성	• 인공지능이 생성한 콘텐츠에 대한 책임 소재와 권리 귀속이 명확하지 않음 • 학습 데이터의 출처와 생성물 활용의 적법성에 대한 기준이 부족함 • 이용자가 생성한 콘텐츠를 인공지능이 다시 학습하고 다른 이용자의 답변에 활용하는 사례에 대한 이해가 명확하지 않음
이용자 보호 방안 부족	• 생성형 인공지능의 기술적 한계와 윤리적 문제를 감안한 이용자 보호 방안이 미흡함 • 생성형 인공지능 서비스 이용자 보호에 대한 가이드라인이 구체적이지 않으며, 기존 가이드라인과의 차별성을 찾기 어려움

문제	구체적 논의 내용
국내외 차별적 접근의 어려움	• 국내외와 해외의 인증 시스템 차이로, 국내는 본인 확인 제도가 있어 연령 기반 접근 제한이 가능하지만, 해외는 동일한 방식 적용이 어려움
문구 조정	• ‘이용자를 보호합니다’라는 표현을 ‘보호하기 위한 수단이나 조치를 마련해야 합니다’로 완화할 필요가 있음 • 사업자마다 실천사례를 지킬 수 있는 범위가 다르므로, 단정적인 표현 대신 권장사항으로 문구를 조정할 필요가 있음

자료: 연구진 작성

초안에 대한 의견 개진 이후 지적된 문제들을 어떻게 개선할지에 대해 논의하였다. 회의에서 논의된 개선을 위한 의견들은 <표 3-22>에 정리하였다.

<표 3-22> 가이드라인 초안의 문제점을 해결하기 위한 개선방안 논의

개선방안	구체적 논의 내용
분류의 명확화	• 인격권, 프라이버시, 저작권 등 각 항목을 명확히 정의하고 중복을 최소화 • 필터 버블과 편향성 문제를 일관되게 다루고, 표현을 조정할 필요
청소년 유해 콘텐츠와 불법 콘텐츠 구분 강화	• 인공지능 생성 콘텐츠에 관해 필터링 시스템을 강화하여, 성인 인증 및 경고 시스템을 도입하여 유해 콘텐츠에 대한 접근을 제한 • 청소년 유해 검색어 입력 시 제한적인 결과 제공 또는 결과를 제공하지 않는 방식의 도입이 필요 • 청소년 유해 콘텐츠와 불법 콘텐츠 구분에 관해 청소년 보호를 위한 구체적인 가이드라인 마련이 필요(등급제 접근 방식의 과도함을 피하고, 사업자들의 현재 상황을 고려한 조정 필요)
책임 및 권리 범위 명확화	• 인공지능이 생성형 콘텐츠의 권리와 책임 범위를 명시 • 서비스 제공자는 콘텐츠 필터링 및 보호 조치를 통해 인공지능이 유해 정보를 생성할 시 이를 고지하거나 신고 절차를 마련하는 등 안내
이용자 보호 중심의 내용 강화	• 서론에서 본 가이드라인의 차별성, 생성형 인공지능 서비스 이용자 보호의 중요성, 서비스 제공자 강조 필요 • 기존 인공지능 가이드라인과의 차별점을 다루어야 함
국내외 법적 차이를 고려한 대응	• 국내에서는 본인 인증 기반의 연령 제한 시스템을 적극 활용하며, 해외의 경우 OECD의 인공지능 프레임워크를 참고하여 생성형 인공지능의 범위를 명확히 시각적으로 표현하는 국제적으로 통용 가능한 기준을 마련함
문구에 관한 지속적 업데이트와 사업자 의견수렴	• 본 가이드라인은 초기 단계에 불과하므로, 해당 가이드라인 초안을 유동적으로 운영하며, 기술발전과 사회적 변화에 따라 점진적으로 사안을 구체화시켜가며 보완함

개선방안	구체적 논의 내용
문구에 관한 지속적 업데이트와 사업자 의견수렴	• 실제 사업자들이 하고 있는 구체적인 실행사례를 포함해 실천 방안을 강화하는 것이 필요 • 현재 시행되고 있는 실행사례를 통해 실천 방안이 현실적이고 적용 가능한 것임을 보여주어야 함

자료: 연구진 작성

5. 제5차 가이드라인 연구반 전체회의

4차 회의 이후 약 한 달 동안 실무반에서 수정/보완 작업을 수행하였다. 4차 회의에서 나온 의견이 다 반영된 수정안이 마련된 이후에는 5차 회의 전까지 남은 한 달 동안 사업자들을 만나서 의견을 수렴하였고, 사업자 의견을 정리하여 5차 회의 때 집중 논의하였다. 사업자 의견은 다음 3절에서 다룰 예정이며, 여기서는 연구반 회의 내용만 다룬다.

제5차 전체회의가 8월 16일(금) 상연재 서울역점에서 개최되었다. 지난 4차 전체회의에서 가이드라인 초안에 대해 문제점과 개선사항에 대해 논의하였다면, 본 5차 회의에서는 수정된 가이드라인과 그동안 있었던 기업간담회에서 나온 사업자들의 의견을 공유하고 그에 대해 논의하는 시간을 가졌다.

앞서 회의 때 나온 주제 외에 새로운 내용만 간단히 다루면, 먼저 현시점에서의 생성형 인공지능 기술과 서비스는 이제 막 출연한 이머징 기술 및 서비스라는 점을 염두에 두어야 한다는 것이었다. 기술의 빠른 변화뿐만 아니라 어떤 방식으로 비즈니스 모델이 완성될지에 대해 아직 확실한 경험이 없는 상태에서 선불리 확정적인 법령을 만드는 것보다 언제든지 유연하게 수정, 보완할 수 있는 가이드라인 형식이 타당하다는 데 의견을 다시 모았다.

또, 기업의 현 상황과 실태를 파악하고 합리적 방안을 마련해주는 것이 타당하다는 데도 의견을 같이하였다. 예를 들어, 유통 과정과 상품성, 퍼블리시티권의 타당성, 데이터 처리 문제 등에 대해 현재로서는 긍·부정의 판단을 내릴 경험적 근거가 부족하다는 점을 인정하고, 가이드라인에 해당 내용을 담는다고 하더라도 강경한 글투보다는 실사례와 현황을 같이 제시하면서 향후 어떤 방향으로 개선해야 하는지에 대한 지침을 담자는 것이다.

이를 위해 서비스 현황을 파악하기 위해 그동안 수집한 실천사례들을 부록으로 담아내는 의견이 새롭게 나왔다. 각 지침과 관련하여 현재 선발주자 서비스 제공자들이 실천

하고 있는 사례들을 제시하자는 것이다. 만약 구체적인 실천사례가 없다면 국내외 법제 동향에서 현재 규제 논의가 어떻게 진행되고 있는지를 제시하여 지침의 타당성을 높이는 안도 나왔다.

다른 한편으로는 생성형 인공지능 서비스의 특수성을 고려할 때 모델이나 주로 활용한 학습데이터 소스 등을 투명하게 공개하는 것의 중요성, 인권 침해 예방 및 이용자 자율성 보호를 위한 사업자 노력 등은 주요하게 다루어야 한다는 점을 강조하기도 하였다. 그밖에 다양성 평가 기준이나 모니터링, 위험관리 거버넌스 체계 등에 대해서도 논의하였다.

6. 제6차 가이드라인 연구반 전체회의

5차 회의 이후 다시 수정, 보완 작업이 있었으며, 글로벌 기업인 마이크로소프트(MS)와 구글, OpenAI의 의견을 받았다. 이 글로벌 기업은 한글 작업한 가이드라인 초안을 자체적으로 다시 영문화하여 본사에 보내 검토받은 후 다시 한국지사에서 한글화 작업하여 의견을 보내주는 절차를 거쳤다. 이 과정이 대략 3개월 가까이 소요되었다. OpenAI는 한국지사가 없는 대신에 김앤장에서 대리하고 있었다.

글로벌 기업에서 검토하는 동안 국내 기업들의 2차 검토도 병행되었다. 글로벌 기업이 검토하는 버전으로 국내 기업들이 다시 검토할 수 있게 한 것이다. 이때 1차 검토에서 참여하지 않았던 KT가 2차 검토에 참여하였다. 2024년 KT는 생성형 인공지능 서비스 관련하여 BtoC 시장에 진출하지 않았기 때문에 1차 검토에서 빠졌으나, 2025년 이후 BtoC 시장 진출을 염두에 두게 되면서 2차 검토에 참여하게 된 것이다.

글로벌 기업의 과정이 마무리된 후 다시 연구반 전체회의를 하여 결과물에 최종 합의하였다. 이 6차 전체회의는 10월 18일(금) KISDI 3층 소회의실에서 온라인 회의와 병행하여 개최되었다. 이때 합의된 내용은 4절에서 다루어질 가이드라인 최종(안)에 반영되었기에 여기서는 생략한다.

제3절 가이드라인(안)에 대한 사업자 의견수렴

이번 절에서는 가이드라인 개발 과정에서 서면 조사 및 대면 간담회 등의 형식으로 이루어진 사업자 의견수렴 내용을 담았다. 전체 과정에서 사업자 의견수렴은 크게 세 부분으로 나눌 수 있다. 첫째, 3월에 서면 조사 문항을 구성하여 대표 사업자에게 서비스 및 이용자 보호 조치에 대한 현황과 자율규범 방향에 대한 의견을 받았다. 글로벌 기업의 경우 서면 답변에도 본사의 검토 과정이 필요하였기에 답변은 5월 초까지 받았으며, 그 결과는 5월 연구반 회의에서 다루어졌고 초안을 작성하는 데 참고자료로 활용하였다.

둘째, 6월 연구반 전체회의(5차)에서 검토를 마친 초안을 가지고 기업 간담회 형식으로 국내외 사업자를 만나 의견을 들었다. 첫 간담회는 한국인터넷기업협회에서 있었으며, 네이버, 카카오, 그리고 AWS와 메타의 실무진이 참여하였다. AWS와 메타는 국내에서 생성형 인공지능 서비스를 하는 사업자는 아니었지만, 향후 비즈니스에 해당 가이드라인이 미칠 수 있는 영향을 고려하여 참여 의사를 밝혔다. 두 번째는 대표 스타트업을 직접 방문하여 의견을 청취하였는데, 연구진이 오전과 오후에 각각 퀴텐테크놀로지스와 스캐터랩을 직접 방문하였다. 세 번째 간담회는 통신사인 SK텔레콤과 LG유플러스 담당자와 진행하였다. 이때 KT는 BtoC 사업을 진행하지 않고 있다는 이유로 사업자 간담회에서는 제외하였다.

셋째, 마지막 의견수렴은 6월 연구반에서 검토하고 7월-8월 사업자 의견수렴까지 반영한 버전으로 진행하였다. 이때는 대면으로 진행된 사업자 간담회 때는 빠졌던 글로벌 기업 마이크로소프트(MS)와 구글, OpenAI가 참여하였다. 앞서 언급하였듯이 이 글로벌 기업은 본사의 검토가 필요하였기에 서면으로만 진행할 수 있었다. 서면 의견수렴에는 국내 기업들도 다시 참여하였다. 사업자 간담회 때의 버전에서 내용 수정이 많이 이루어졌고, 또 사업자 간담회 때 내놓은 의견이 어떻게 반영되어 있는지 검토할 기회를 주기 위해서였다.

이와 같이 가이드라인 개발 과정 동안 전체 세 시기에 걸쳐 이루어진 사업자 의견수렴 내용을 다음에서 소개한다.

1. 사업자 현황 조사 및 자율규범 방향에 대한 의견수렴

가이드라인을 본격적으로 개발하기에 앞서 3월에 국내 생성형 인공지능 서비스 대표 사업자들에게 서비스 및 이용자 보호 조치에 대한 현황과 자율규범 방향에 대한 서면 의견을 구하였다. 이 서면 조사는 연구진이 사전에 조사한 서비스 현황을 사업자들로부터 다시 확인하고 심화된 자료를 받기 위한 목적과 함께, 생성형 인공지능 서비스 이용자 보호 가이드라인을 개발할 것임을 알리고 협조를 구할 목적으로 수행되었다. 참여 사업자는 국내 대표 플랫폼기업인 네이버와 카카오, 통신사 SK텔레콤과 LG유플러스, 스타트업 쿠팡테크놀로지스와 스캐터랩, 그리고 글로벌 기업인 구글, MS, OpenAI로 총 9개 기업이 참여하였다. 서비스 및 이용자 보호 조치에 대한 주요 문항은 <표 3-23>과 같다. 각 문항에서 필요에 따라 증빙자료를 첨부할 것을 요청하였다. 가이드라인 개발 시 염두에 두어야 할 것에 대한 의견은 따로 담지 않을 예정이며, 대략 사업자별 이용자 보호 조치에 대한 상세한 설명과 함께, 개발하는 가이드라인이 사업자 규제에 영향을 미치는 것에 대한 우려를 표명하였음을 밝힌다. 이 서면조사를 통해 받은 의견은 초안을 작성하는 데 참고자료로 활용하였으나, 상반기 사업자 현황이 하반기와 다른 경우도 많아서 계속 수정, 보완하여야 했다.

<표 3-23> 사업자 의견수렴: 서비스 및 이용자 보호 조치에 대한 현황 조사

주제	문항
환각현상	Q1-1. 환각 현상 관련 고지 방법 Q1-2. 고지 문구
투명성·설명가능성	Q2. 생성 콘텐츠 오류 관련 이용자 판단 돕는 검증방법 Q3. 생성형 AI 데이터 구성 관련 정보 공개 여부
학습데이터 활용 관련 고지 및 이용자 동의	Q4-1. 생성형 AI 학습을 위해 이용자의 학습데이터 활용 여부 Q4-2. 이용자에게 AI 학습데이터 사용 사실 고지방법 Q5. 본인 데이터의 학습데이터 사용에 대한 동의 절차
학습데이터 활용 거부권	Q6. 이용자가 본인 데이터의 학습데이터 거부 선택 가능 여부
입력데이터 저장·수집	Q7. 이용자 데이터 저장 기간 Q8. 이용자 데이터 저장, 수집 이유

주제	문항
이용자 피해발생시 요청 절차	Q9-1. 인공지능 생성 콘텐츠로 인한 권리 침해 시 절차 Q9-2. 요청할 수 있는 구체적 상황 제시 여부
인공지능 편향성	Q10. 알고리즘 편향성 대응책
인공지능 생성 콘텐츠 오남용 방지	Q11-1. 이용자의 오남용 대응책 Q11-2. 사용하고 있는 생성물 표시 방법 Q11-3. 생성물 표시 변조·삭제를 막기 위한 기술적 조치 Q11-4. 기술적 탐지와 그 밖의 대응방안 Q12. 개발자의 통제범위를 벗어난 탈옥에 대한 대응책
미성년자·청소년 보호 노력	Q13. 미성년자 유해 콘텐츠 생성 우려에 대한 대응책 Q14. 미성년자·청소년 이용자의 보호를 위한 구체적인 노력
인공지능 위험관리	Q15. 인공지능 위험성을 예방하기 위한 내부 통제 체계

자료: 연구진 작성

해당 문항들에 대한 사업자 답변을 대략 소개하면 다음과 같다. 먼저 인공지능 생성 콘텐츠 오남용 방지와 관련하여, 한 기업은 자체 개발한 SynthID를 통해 AI 생성물에 비가시적 표시를 추가하고, 데이터 입력 및 출력 단계에서 권리 침해를 방지하는 필터링 기술을 적용하였다. 또 응답 기업들은 레드팀 프로그램을 운영하여 인공지능을 통한 사이버 보안의 취약성을 분석하고 이상 신호 탐지, 인공지능 생성 콘텐츠의 진위 검증 등의 기술을 갖추고 있었다. 인공지능 생성물에 워터마크를 삽입하여 콘텐츠가 인공지능에 의해 생성되었음을 명시하거나, 이미지와 문서의 필터링 기술을 도입하여 잠재적 인권 침해를 방지하고자 하였다. 생성형 인공지능 콘텐츠에 사용된 원저작물을 표기하고 고지하는 방안을 계획 중인 기업도 있었고, 개인화 서비스를 위해 고객 데이터를 활용할 시 민감 정보를 완전히 가명 처리하여 오남용 가능성을 차단하는 절차를 필수로 하였다. 인공지능 생성 콘텐츠의 워터마킹 및 태깅에 대한 독립적인 사양을 개발하는 조직, C2PA(Coalition for Content Provenance and Authenticity)를 통해 인공지능 생성물의 신뢰성을 보장하고, 생성물 필터링과 검증 기술을 포괄적으로 적용하는 사례도 소개하였다.

다음은 환각 현상 조치에 대한 답변이다. 모든 기업에서 인공지능 생성 콘텐츠가 사실이 아닌 정보를 제공할 수 있음을 알리는 고지 문구를 두고 있었다. 먼저 서비스 약관 및 FAQ와 고객센터에서도 생성형 인공지능 서비스의 한계와 부정확성 및 관련 정보를 제공하는 형태가 있었다. 또 실행화면을 통해 데이터 활용 및 콘텐츠 신뢰성에 대한 정보를 사용자

에게 안내하는 사례, 약관과 실행화면에서 인공지능 생성 콘텐츠의 신뢰성을 사용자 스스로 재확인할 수 있도록 안내하는 사례 등이 있었다. 여러 기업이 이용약관, 실행화면, 고객센터를 통해 생성형 인공지능의 오류 발생 가능성을 다각도로 고지하였으며, 어떤 기업은 서비스 약관과 고객센터에서 인공지능 생성 콘텐츠의 한계와 책임을 명시하였다. 또 서비스에서 인공지능 생성 콘텐츠의 출처를 명확히 표기하고 검증 가능한 외부 링크를 제공함으로써 투명성을 확보하는 경우, 그리고 실행화면과 FAQ를 통해 인공지능 생성 콘텐츠의 부정확성을 사용자에게 안내하는 경우 등이 있었다.

이용자의 입력 자료를 인공지능 학습데이터로 활용한다고 응답한 기업 중에 학습데이터 수집 및 활용을 이용자에게 어떤 방식으로 고지하는지에 대해서는 다음과 같은 경우를 확인하였다. 최초 서비스 실행 시 사용자 동의를 명시적으로 받거나, 입력창과 팝업 화면에서 데이터 학습 활용 여부를 고지하는 경우였다. 또 수집되는 텍스트/음성/사진 정보를 기능 향상을 위해서만 사용한다는 점을 서비스 약관에 밝히거나 데이터 가명처리와 민감 정보 삭제로 안전한 학습데이터 활용을 보장하는 사례도 있었으며, 대부분 이용약관에 학습데이터 활용 방식을 명시하여 이용자에게 관련 정보를 제공하였다. 일부 기업은 이용자가 데이터 학습 사용 여부를 선택할 수 있도록 옵션을 제공하였다.

콘텐츠 유통에서 발생하는 책임 소재 문제에 관해 각 기업에서 인공지능 위험성을 예방하기 위한 내부 통제에 대해서도 답변을 얻었다. 다수 기업이 인공지능 위험성을 분석하고 잠재적 문제를 사전 탐지하기 위해 레드팀을 운영하였으며, 안전성 확보를 위해 정책을 수립하였다. 또, 자체 인공지능 윤리기준을 개발하고 서비스의 안전성을 강화하는 경우도 다수였으며, 기존 보안 체계 내에서 인공지능의 위험성을 관리하는 대응방안을 따로 두기도 하였다. A사는 문제 발화를 탐지하고 내부 비상대응 체계를 운영하여 인공지능 오작동에 대응한 경우를, B사는 잠재적 문제를 사전에 평가하고 안전성을 강화한 사례를 소개하였다. C사는 외부 전문가와 협력하여 인공지능의 위험성을 분석하고 개선한 사례를, D사는 인공지능 취약성 분석과 잠재적인 인공지능 위험 식별 및 해결을 위한 레드팀을 운영하고 인공지능 위험관리 정책을 마련한 과정을 적었다.

불법 콘텐츠와 유해 정보 노출 관련하여 각 기업은 미성년자 및 청소년 보호 조치를 취하고 있었다. 예를 들어, 연령 제한과 유해 콘텐츠 차단 기술을 도입하여 불법 콘텐츠 및 유해 정보가 노출될 수 있는 환경으로부터 청소년을 보호하였고, 사진 같은 입력물에 대

해 적합성 확인 후, 노출이 심한 사진의 경우 부적절한 사진으로 안내하는 단계도 있었다. 회원 가입 시 연령 제한을 적용하는 경우가 다수였고, 일부 서비스만 청소년에게 허용하거나 유해 콘텐츠를 방지하는 필터링 단계를 계층화하여 청소년이 유해 콘텐츠에 노출되는 경우의 수를 줄이려고 하기도 하였다. 선정적 콘텐츠를 탐지하는 탐지 기술을 사용하거나 위반 사용자에게 페널티를 부과하는 등의 조치도 있었다. 어떤 기업은 소위 ‘엄격모드’를 기본 설정으로 하여 미성년자가 유해 콘텐츠에 노출되지 않도록 하였고, 서비스 가입 정책으로 13세, 14세 이하는 서비스 가입을 제한하거나 19세 이상의 실명 회원에게만 가입을 허용하기도 하였다.

생성형 인공지능의 편향성 이슈에 대한 대응책은 다음과 같았다. 편향 콘텐츠를 탐지하고 차단하는 필터링 기술을 도입하거나 내부 윤리 위원회를 구성하여 기계학습 공정성을 검토하는 것이다. 또 인공지능 모델에서 편향성을 최소화하기 위해 안전 필터링을 적용하고, 이용자에게 편향성 콘텐츠에 대한 신고를 받는 시스템도 마련하였다. 그 밖에 자체 단어 필터링 기술과 외부 도구를 활용하여 윤리적 문제를 방지하거나, 어뷰징 탐지 모델을 도입하여 부적절한 발화를 감지하는 사례도 있었다. 위험관리 거버넌스를 따로 운영하거나 외부 전문가와 협력하여 편향성 문제를 분석하고 개선하는 사례도 있었다.

2. 초안에 대한 사업자 간담회

사업자 간담회는 6월 연구반 전체회의(5차)에서 검토를 마친 초안을 가지고 이루어졌으며, 크게 세 차례로 구분할 수 있다. 첫 간담회는 한국인터넷기업협회에서 있었으며, 네이버, 카카오, 그리고 AWS와 메타의 실무진이 참여하였다. AWS와 메타는 국내에서 생성형 인공지능 서비스를 하는 사업자는 아니었지만, 향후 비즈니스에 해당 가이드라인이 미칠 수 있는 영향을 고려하여 참여 의사를 밝혔다. 두 번째는 대표 스타트업을 직접 방문하여 의견을 청취하였는데, 연구진이 오전과 오후에 각각 뽀빠테크놀로지와 스캐터랩을 직접 방문하였다. 세 번째 간담회는 통신사인 SK텔레콤과 LG유플러스 담당자와 진행하였다. 이때 KT는 BtoC 사업을 진행하지 않고 있다는 이유로 사업자 간담회에서는 제외하였다.

〈표 3-24〉 사업자 대면 간담회 일정

일정	장소	참여기업
2024. 7. 16.	한국인터넷기업협회 회의실	네이버, 카카오, 메타, AWS
2024. 7. 31.	뤼튼 테크놀로지스 회의실 스캐터랩 회의실	뤼튼 테크놀로지스, 스캐터랩
2024. 8. 8.	양재 토즈	SK텔레콤, LG유플러스

가. 첫 번째 간담회: 네이버, 카카오, AWS, 메타

첫 간담회는 4개 기업이 참여하였다. 먼저 참여 기업들이 의견을 같이한 것으로, 본 가이드라인의 자율 규제적 특성을 명확히 해달라고 요청이 있었다. 또 원칙 및 의미가 명확히 전달되기를 바랐는데, 예를 들어 실천방안에 주어를 분명히 하여서 수범자를 확대하지 않기를 원하였다. 즉, 개발사에만 해당하는지, 서비스 제공자만 해당하는지, 둘 다 해당하는지 등을 명확히 해달라는 요청이었다. 또, 실천방안 중 이용자의 프롬프트 입력값에 대한 검토와 관리 항목 관련하여 그러한 조치가 오히려 이용자 프라이버시와 개인정보 노출 위험성을 수반할 수 있으며, 따라서 기업이 이를 완전하게 관리하는 것이 현실적으로 어렵다는 의견을 개진하였다.

가이드라인의 필요성과 관련해서는 생성형 인공지능의 위험성에만 초점을 맞추기보다 올바른 활용을 통한 긍정적인 결과와 영향력을 제시해 달라는 요청도 있었다. 자칫 이용자가 생성형 인공지능 서비스의 유용성을 경험하기 전에 벌어지지 않은 잠재적 위험으로 서비스 이용에 저항감을 가질 수 있음을 염려한 것이다. 또 현시점의 생성형 인공지능 서비스는 이미 학습이 완료된 인공지능 모델을 사용하기에 이용자가 인공지능과 상호작용하면서 생성된 콘텐츠를 바로 학습데이터로 사용하지 않는다는 점도 논의되었다. 그러기에 학습데이터 관리에 대한 항목이 빠져야 한다는 의견이었는데 이에 대해서는 연구진과 의견 차이가 있었다. 지금은 아니더라도 인공지능 에이전트가 보편화하는 시점에서 이용자-인공지능 상호작용 과정이 모두 학습되는 개별화 서비스가 많아질 것이므로 빠져서는 안 된다는 것이 연구진의 의견이었다. 추가로 학습을 원하지 않는 이용자의 데이터만 제외하는 것이 현실적으로 불가능하다는 의견도 나왔다.

서비스 오용에 대한 책임은 최종 이용자가 지는 것이 타당하며, 개발자와 배포자는 오

용을 완화하기 위한 기술적, 정책적 조치 의무에 한정해야 한다는 의견도 있었다. 또 개발자가 모든 소비자의 피해를 탐지하고 통제하는 데 기술적 한계가 있음을 고려해야 한다거나 ‘필터버블’이라는 표현이 생성형 인공지능의 특성과 맞지 않으며, 이용자가 지속적으로 특정 결과물을 요구한 결과로 설명하는 것이 더 적절하다는 의견도 있었다.

〈표 3-25〉 1차 사업자 간담회 주요 내용(네이버, 카카오, AWS, 메타 참여)

기업	주요 내용
A	• 이용자의 프롬프트에 대한 개입은 이용자의 프라이버시 문제와 개인정보 노출 등 오히려 위험성을 수반하는 상황이 될 수 있으며 현실적으로 기업에서 통제가 어려운 영역임 • 개인정보 사용에 대한 명확한 고지와 투명한 설명이 필요하며, 이와 관련한 논의는 가이드라인 내 별도 구성으로 다루어야 함
B	• 가이드라인의 필요성 부분에서 부정적인 영향이 강조된 경향이 있으므로, 올바른 활용을 통한 긍정적인 영향을 제시하는 문단이 필요 • 필터버블 현상은 추천 시스템에서 부각되는 문제로, 이용자의 요구에 따라 결과물을 만들어내는 생성형 인공지능의 문제라기보다 특정 응답을 원하는 이용자 행동을 고려해야 함 • ‘사전 모니터링’이라는 용어는 이용자의 입력 데이터, 정보에 접근할 수 있는 것처럼 묘사되고 있기에 오해의 소지를 일으킬 수 있어, 수정이 필요함
C	• 실천방안 속 제시된 로그데이터는 인공지능의 학습에 쓰이지 않으므로, 오히려 혼란을 야기하거나 추가적인 관리 비용을 발생시킬 수 있음 • 이용자의 입력 데이터를 다시 학습하지 않으므로, 이에 대한 관리가 불필요함 • 기존 학습에 이용된 데이터는 개별 이용자가 원치 않더라도 따로 추적하여 제외하는 것이 불가능함
D	• 인공지능 관련 규제에서 파운데이션 모델 LLM까지 모두 규제의 대상으로 바라보는지, 소비자와 맞닿는 부분만 규제하는 것인지 명확하게 정리될 필요가 있음 • 4, 5, 6번 대주제에 사용된 ‘사전 모니터링, 프롬프트 입력에 대한 필터링’ 등의 용어는 이용자의 입력에 대해 개발사나 서비스사가 개입할 수 있는 것처럼 느껴질 수 있어 조정이 필요함

자료: 연구진 작성

나. 두 번째 간담회: 뤼튼 테크놀로지스, 스캐터랩

생성형 인공지능 분야 국내 대표 스타트업인 뤼튼 테크놀로지스와 스캐터랩은 연구진이 각 기업을 방문하여 간담회를 진행하였다. 간담회에서 나왔던 의견들을 대략 정리하면, 먼

저 생성형 인공지능 관련 사항을 법안으로 제정하기에는 시기상조라는 의견뿐만 아니라, 법령 제정이 아니라 ‘지침서’ 형태로 나오는 데 대한 긍정적 의견이 있었다. 생성형 인공지능 서비스로 인한 사회 이슈가 발생한 만큼 어느 정도의 사회적 통제 가능성을 증명하여 이용자의 불필요한 저항감을 완화할 필요가 있다는 것이다. 또 다른 의견으로는, 생성형 인공지능 모델을 명시하는 데 대한 어려움을 호소하였다. 2024년 상반기까지만 해도 특정 모델을 중심으로 서비스가 론칭되었지만 현시점에서는 시중의 여러 모델을 복합적으로 사용하고 있기에 모델 정보 제공의 의미가 퇴색했다는 것이다. 토론 끝에, 모델 정보 제공은 사업자의 자율에 맡기는 것이 최선이라는데 의견을 같이하기도 하였다. 투명한 정보 제공은 서비스 신뢰를 향상하는데 이점이 되지만, 복잡한 설명은 오히려 이용자의 외면을 받을 수 있기 때문이다.

인공지능에 대한 정의가 주체마다 다르게 이해되고 있어 이를 단순화하거나 기존 법의 틀 안에서 설명할 수 있도록 최소화하는 것이 현실적이라는 점도 논의되었고, 관련하여 영향력 있는 법안에 근거하여 정의해 줄 것을 요청받았다. 편향성 이슈에 대해서도 논의되었는데, 편향성 관련하여 사회적으로 합의된 가치 기준이 부족하다는 문제가 제기되었다. 현재 데이터에 내재한 편향성을 제거하기 위해 튜닝 과정을 거치는데, 수많은 경우를 판단하는 기준이 부족하다고 하였다. 하나의 이슈에 여러 가치가 충돌하는 경우가 자주 일어나고, 그러한 경우 매우 보수적으로 알고리즘을 처리하는 데 이 경우 고객의 니즈를 충족시키지 못하는 상황이 발생하는 딜레마가 있다는 것이다. 또, ‘필터버블’이라는 용어는 생성형 인공지능에서 적합하지 않을 수 있으며, 편향성 문제에 초점을 맞추는 방향으로 조정할 필요성이 제기되었다.

조직의 규모가 크지 않고 자금이 충분하지 않은 스타트업의 특수성을 고려해야 한다는 의견도 제기되었다. 위험관리를 위해 별도의 조직이나 팀을 구성하고 운영할만한 규모가 아니라는 것이다. 제3자 기관에서 위험관리를 외주하는 것에 대해서도 부정적이었다. 생성형 인공지능 서비스 산업은 이제 막 비즈니스 모델을 정립해가는 초기 단계이므로 기업의 정보가 새어나갈 수 있는 경로를 최소화하고 싶다는 것이다. 대안으로 책임자를 지정하는 안을 제안받았는데, 구성원 모두가 테스트와 위험관리에 참여하는 소규모 스타트업에게는 책임자 지정이 가장 실현 가능한 선택지라고 하였다.

〈표 3-26〉 2차 사업자 간담회 주요 내용(뤼튼 테크놀로지스, 스캐터랩)

기업	주요 내용
E	• 현실적이고 풍부한 시나리오가 생성형 인공지능 서비스의 효과를 높이고 실질적인 도움을 줄 수 있기 때문에 실질적이고 다양한 실행사례 제공 필요 • 생성형 인공지능은 엔터테인먼트 분야에서 활성화되며, 고객의 니즈에 반하거나 품질 저하를 초래하는 규제는 국내 서비스 실패로 이어질 수 있음 • 보수적인 알고리즘 필터링으로 오히려 대중에게 필요한 정보를 제공하지 못하는 경우가 발생함. 예를 들어, 인물 사진 입력 제한은 정보를 얻고자 하는 고객의 니즈를 충족시키지 못하고 정보의 공공성도 저하시킬 수 있음 • 생성물에 대해 워터마크를 남기는 방식은 기업의 비용 부담 대비 효과가 낮음. 콘텐츠 품질 저하 및 고객 불만 초래, 나아가 해당 규제가 없는 해외의 서비스로 이용자들이 이주하는 현상을 발생시켜 국내 인공지능 서비스 생태계에 악영향을 미칠 수 있음 • 생성형 인공지능 모델 정보 제공에 관해, 현재는 복합 모델 사용이 일반적이기 때문에 오히려 정보 제공이 사용자 혼란 및 악성 유해 공격을 유발할 수 있으며, 최신 모델 안내는 빠른 교체 주기로 인해 기업에 부담이 됨 • 학습 데이터 관리 및 개인정보보호의 내용은 타 정부 부처의 안내와 중복되어 적용되는 부분이 있고, 이에 대해 서로가 제시하는 바를 동시에 충족시킬 수 없는 지점이 있음. 이러한 지점에 대해 정부 부처 간 협업이나 논의가 이루어지길 희망함
F	• 인공지능에 대한 정의를 최소화하거나 기존 법에 따라 현실적으로 접근해야 할 필요가 있으며, 개인정보보호위원회의 정책과 충돌하지 않도록 조정하여 사업자에게 과도한 부담을 주지 않도록 해야 함 • 대법원 판례나 입법으로 확립되지 않은 퍼블리시티권을 선제적으로 적용하는 대신, 유명인의 권리로 우회 표현하는 방안을 고려해야 함 • 가이드라인에서 부정적 어휘(통제) 대신 완화된 표현(관리)을 사용하고, 개인정보 입력과 혐담의 예시는 중요도를 구분하여 조정해야 함 • 생성형 인공지능은 추천 알고리즘과 달리 개인화가 제한적이므로, 필터버블 문제가 가이드라인에서 논의된 것처럼 현실화되지 않을 수 있다는 우려가 있음 • 학습데이터 관련 모든 정보 공개는 산업계 반발을 초래할 수 있으므로, 설명 가능성을 보완하는 선에서 적절한 공개 범위를 설정해야 함 • 스타트업의 조직/인력 규모를 고려하여 위험관리 책임의 경우 별도 조직이나 팀 외에 책임자 지정도 포함할 것을 제안함

자료: 연구진 작성

다. 세 번째 간담회: SK텔레콤, LG유플러스

세 번째 간담회는 국내 이동통신사 SK텔레콤과 LG유플러스가 참여하였다. 이들은 본 가이드라인의 취지와 목적에 대해 전반적으로 공감하면서, 생성형 인공지능 서비스의 유

형별로 가이드라인이 제정되었으면 하는 의견을 제기하였다. 또, 이미 주요 내용은 기업 내부정책과 거버넌스에 반영되고 있다고 하였으며, 국내와 국외 사업자 간 역차별 이슈를 방지하기 위한 노력의 필요성을 강조하였다.

가이드라인은 사업자 준수사항과 이용자 준수사항을 구분하여 주체별로 명확히 안내해야 하는 것이 중요한데, 인공지능 모델에 대한 정보 제공보다는 이용자의 질의에 대해 이해하기 쉽고 신뢰할 수 있는 답변을 제공하는 방식이 투명성을 높이는 데 효과적이라고 언급하였다. 일부 서비스에서는 다양성 보장을 위해 편향성을 최소화할 필요가 있으나, 개인화된 서비스의 경우 이용자의 선호도를 반영한 맞춤형 결과 제공이 중요하며, 다양성과 개인 맞춤화 간 균형을 유지해야 한다는 점을 강조하였다.

생성형 인공지능 서비스 제공자는 이용자가 서비스 활용 과정에서 발생할 수 있는 문제에 대해 경고 메시지나 안내를 통해 기본적인 주의의무를 다하고 있으나, 악의적이거나 불법적인 활용을 완전히 통제하기는 어렵다는 점을 강조하기도 하였다. 또, 이용자 피해를 줄이기 위한 모니터링 의무화는 인력과 비용 부담으로 이어질 수 있으며, 특히 스타트업에게는 과도한 부담이 될 수 있다는 의견을 주었다. 위협관리체계와 관련해서는, 이용자 피해 발생 최소화를 위해 기업 내 위협관리 체계를 갖추는 것은 필요하지만 사전 모니터링이나 상시 감독과 같은 표현은 실효성이 없다는 의견이었다. 사전 모니터링은 ‘감시’와 연결될 수 있고, 또 ‘상시’라는 용어가 주는 기대감, 즉 24시간 매초의 모니터링은 비현실적이라는 것이다. 이에 용어의 완화를 제안하였다.

〈표 3-27〉 통신사 사업자 의견 수렴 주요 내용

기업	주요 내용
G	• 생성형 인공지능 서비스 유형에 따른 가이드라인 제정을 고려할 필요가 있으며, 국내 사업자와 해외 사업자 간의 역차별 이슈가 발생하지 않도록 노력을 기울일 필요가 있음 • 이용자가 생성형 인공지능의 결정 과정을 이해할 수 있도록 인공지능 모델에 대한 정보를 제공하기보다는 이용자의 질의에 이해하기 쉽고 신뢰할 수 있는 답변을 제공하는 방식이 투명성 제고 차원에서 오히려 유용할 수 있음 • 일부 서비스에서는 다양성 보장을 위해 편향성을 최소화할 필요가 있지만, 개인화된 서비스의 경우에는 이용자의 선호도에 따라 개인 맞춤형 결과가 제공될 필요도 있으므로 다양성과 개인 맞춤화 사이의 균형을 맞추는 것이 중요함

기업	주요 내용
G	• 생성형 인공지능 서비스 제공자는 이용자에게 생성형 인공지능 서비스를 활용하는 과정에서 발생할 수 있는 문제에 대해 경고 메시지, 안내, 이용약관 등을 통해 기본적인 주의의무를 다하고 있지만, 이용자의 악의적 또는 불법적인 활용까지 통제하기는 어려움
H	• 이용자가 얼굴 이미지 등 인권을 침해할 가능성이 큰 콘텐츠를 입력데이터로 사용할 때, 사전에 이용자에게 인권 침해의 가능성이나 위험성 등을 안내 또는 경고하는 것은 가능하지만 이를 실시간으로 통제 또는 제한하는 것은 현실적으로 불가능함 • 생성형 인공지능 산출물이 생성형 인공지능으로 생성되었음을 워터마크 등을 통해 표시하는 것은 기술적으로 가능하지만, 이용자가 캡처 등을 통해 워터마크를 제거하여 이를 디지털 환경에 유통하는 경우까지 고려한다면 과연 효과적인 방식인지에 대한 고민이 필요함

자료: 연구진 작성

이상 크게 세 번으로 구분되어 수행된 사업자 간담회에서 공통으로 나온 의견은 주로 기술적 한계와 기준의 모호성에 대한 것이었다. 현재 이용의 오남용을 방지하고자 실행하고 있는 워터마크나 모니터링, 감시시스템 등은 시범적인 수준이지 문제를 완벽하게 해결할 수 있는 수준이 아니라는 것이며, 또 알고리즘 튜닝으로 해결하려는 공정성/편향성 기준 역시 사회적으로 합의된 기준이 모호하여 모두가 만족할 결과물을 생성하기 어렵다는 것이다. 사업자들의 의견을 수렴하고 애로점을 같이 논의하면서 최소한 현시점에서 할 수 있는 최선의 방향에서 이용자 보호에 효과적이면서 실천가능한 가이드라인 개발의 방향을 조정할 수 있었다.

3. 수정안에 대한 사업자 의견수렴

마지막 사업자 의견수렴은 서면 조사로 수행되었다. 7월-8월 사업자 간담회를 통해 받은 의견을 반영하여 일부 조정이 있었고, 대면 간담회에 참여하지 못한 글로벌 기업들에게 조정된 버전을 보내서 검토를 요청하였다. 참여 기업은 MS, 구글, OpenAI이었다. 검토는 한국지사서 본사로 보내서 검토하는 절차가 필요했기에 9월-10월동안 진행되었으며, 한국지사가 없는 OpenAI의 경우는 김앤장이 대리하여 본사 검토를 거친 의견서를 보내왔다. 글로벌 기업들이 수정된 버전을 검토하는 동안, 같은 버전을 대면 사업자 간담회에 참여한 국내 기업들에게도 보내서 다시 한 번 최종 검토할 수 있도록 하였으며, 이때 일련의

과정에 참여하지 않았던 KT도 참여하였다. KT는 2025년 이후 BtoC 시장에 참여할 계획을 세웠기에 검토하고 싶다는 의견을 주었고 이에 수정 버전을 보내 검토하게 하였다. 그 내용은 다음 표에 간략히 정리하였다. 이들의 검토 의견은 기술 표현이나 의견의 초점이 약간 달라도 전체적 맥락은 이미 사업자 간담회에서 나왔던 의견과 크게 다르지 않았으며, 세부 수정 단어/표현 수정 의견은 생략하였다.

〈표 3-28〉 글로벌 사업자*를 포함하여 사업자 최종 의견수렴 주요 내용

주요 내용
• 생성형 AI 개발자 외에도 AI 밸류체인 전체 및 최종 사용자들의 역할을 고려하는 것도 중요함. 생성형 AI 기술의 개발, 배포 및 사용에 있어서 각 이해관계자의 책임을 구별하여 제시할 것을 제안함 • 가이드라인 논의안과 미 국립표준기술연구소(National Institute of Standards and Technology)의 AI 위험 관리 프레임워크(AI Risk Management Framework) 등 국제 이니셔티브 및 표준 기구들에서 채택하고 있는 접근 방식과 가능한 일관되게 유지해줄 것을 요청함. 서로 다른 관할권에서 발생할 수 있는 불일치를 방지할 수 있으며, 개발자 및 배포자들이 일일이 규제 프레임워크 및 지침들에 대응하기 위해 시간을 소요하기 보다 높은 품질의 위험 완화 조치들 개발에 집중하는 데 도움이 됨 • 파운데이션 모델의 범용성으로 인해 AI 개발자들이 모든 소비자 피해를 탐지 및 통제하는 데 기술적 한계가 있음을 고려할 것을 요청함 • 이용자들의 프라이버시를 침해하는 응답을 거부하거나, 차별적 또는 편향적 응답을 방지하기 위한 조치를 취하는 등 이용자 인권권 침해를 제한하기 위한 기존의 모범사례를 활용할 것을 권장하는 내용을 추가하는 것을 제안함 • 개발사가 모든 침해 사례를 방지하는 기술 솔루션을 단독으로 제공해야 한다는 것은 현실적이지 않음을 고려해주길 바람 • 생성형 AI가 어떻게 응답을 생성하는지에 대해 이용자들에게 명확히 설명하고, 응답이 부정확할 수 있는 위험성 등 이 기술의 능력 및 한계를 설명해야 한다는 데 동의함 • 개인정보 처리방침 및 기타 설명자료에 모델 학습 등 개인정보의 이용과 관련한 사항을 명확히 공개하는 것이 중요하며, 이와 관련한 논의는 다른 맥락에서 보다 적절하게 다루어질 수 있으므로, 본 주제2에서는 제외하는 것을 제안함 • 주제3은 다양성과 편향성 감소에 초점을 맞추고 있지만, 모든 생성형 AI 모델이 검색 기록, 브라우징 기록 또는 관심사 데이터에 접근하거나 이용하지는 않다는 것을 고려할 필요가 있음. 따라서 편향성을 감소시키기 위해서는 모델 학습데이터에서 편향된 콘텐츠 양을 감소시키는 것 즉, 응답에 대한 편향성 테스트 및 파인튜닝에 보다 중점을 둘 필요가 있다고 봄 • 생성형 AI는 모델의 정확성을 향상시키기 위해 노력하고 있지만, 애초에 결코 완벽하게 진실되고 신뢰성이 있는 출처가 되거나 응답의 완전한 정확성을 보장하려 한 바가 없음. 따라서, 이 부분의 내용을 개발자 및 서비스 제공자들이 생성형 AI 산출물이 정확하지 않을 수

주요 내용

있다는 사실을 명확하고 투명하게 공개하도록 권장하는 데 초점을 둘 수 있도록 수정하는 것이 필요함

- 이용자의 오용 및 남용에 대한 법적 책임이 AI 시스템의 개발자 및 배포자들에게 있음을 시사하고 있지만, 개발자 및 배포자는 잠재적인 서비스 오용을 감소시키기 위한 조치만 취할 뿐, 그러한 오용에 대한 책임을 부담하지는 않음
- 생성형 AI 기술의 안전한 배포에 있어서는 전체 밸류 체인에서의 각 행위자의 역할을 구분하는 것이 중요하며, 개발자 및 배포자는 기술적 안전 조치 및 정책을 통하여 잠재적 오용의 위험성을 완화할 수 있으나, 경우에 따라서는 최종 사용자들이 이러한 조치들을 무시하거나 우회할 수 있다는 점을 가정하고, 그 경우 생성형 AI 서비스 오용 또는 남용에 따른 책임은 해당 최종 사용자가 부담하는 것으로 해야 함

* 구글, MS, OpenAI

자료: 연구진 작성

제 4 절 가이드라인 주요 주제 구성안

이번 절에서는 생성형 인공지능 서비스 이용자 보호 가이드라인의 주요 주제 구성안을 소개한다. 본 가이드라인은 크게 6개의 주제를 다룬다. 이 6개 주제는 생성형 인공지능 서비스로 인해 발생가능한 이슈 중 저작권과 개인정보 관련 이슈를 제외하고, 인격권 침해, 결정과정의 불투명성, 다양성 침해, 학습데이터 수집·활용 과정에서의 문제, 책임 소재의 불분명, 콘텐츠 유통에서의 문제 등으로 구성되었다. 저작권과 개인정보 이슈를 제외한 이슈는 각각 문화체육관광부와 개인정보보호위원회가 따로 관련 가이드라인을 준비하고 있었기 때문이다.

언급하였듯이 6개의 주제는 관련 이슈들, 즉 이미 사회적으로 이슈화된 문제들에서 출발하였으며, 이 문제들을 어떻게 해소할 수 있을지를 고찰하고 그 방안을 가이드라인의 주요 내용으로 재구성하고자 하였다. 따라서 다음의 본문은 주제별 문제와 해결방안으로 구분하여 기술하였다. 참고로 본 내용은 사업자 의견수렴을 거쳐 최종 수정안이 나오기 전에 작성된 내용이다.

1. 주제1 이용자 인격권 침해

가. 문제점

생성형 인공지능을 기반으로 한 시스템과 광범위하고도 상세한 데이터 분석으로 이용자의 생활에 밀착한 섬세한 서비스를 제공하는 것이 가능해졌다. 그러나 그 결과물의 밀착성으로 인하여 거꾸로 이용자의 프라이버시를 비롯한 인격권을 침해할 수 있다는 위험성 역시 제기되었다.

대화형 서비스 등을 통해 특정 이용자의 개인정보가 노출된다거나, 이용자가 원치 않는 혐오 발언이 표출되는 사례가 일부 서비스에 나타난 바 있다(이효석, 2021). 또 피싱을 통해 이용자의 민감 정보를 수집하거나 재산상의 피해를 주는 결과를 초래할 수 있다. 생성형 인공지능 기반의 검색 서비스는 어떤 내용이든 상세하게 설명해 줄 수 있다는 장점이 있지만, 정교한 데이터의 조합을 통해 오히려 거짓 정보에 해당하거나 편향적이거나 타인의 명예를 훼손하는 결과물을 만들어 낼 수도 있을 것이다. 시스템의 섬세함을 이용하여

특정인을 연상케 하는 사진이나 영상을 만들어 낸다면 소위 퍼블리시티권 침해의 문제로도 이어질 가능성도 있다.

생성형 인공지능의 알고리즘을 통해 만들어진 표현물이 명예훼손 등의 결과를 가져온다 해도, 행위자가 특정되지 않으므로 처벌을 통한 해결이 어렵다는 문제에 직면한다. 개발사 혹은 서비스 제공자가 대개는 생성형 인공지능 기반 시스템을 악의적으로 구성하지 않았을 것이므로, 그들에게 책임을 묻기도 어려운 일이다. 생성형 인공지능이 만들어 낸 정보는 다시 학습데이터로 사용되어, 명예훼손 등의 반복적인 재생산이 가능하기 때문에, 사전 단계에서 이를 관리할 수 있는 방법을 강구하는 일이 중요할 것이다.

나. 해소 방안

생성형 인공지능 기반의 시스템에서 이용자의 인격권을 침해하는 요소를 발견할 수 있게 하고, 관련 요소가 있을 때 시스템 통제를 가능케 하도록 알고리즘을 구축할 책임을 개발자에게 부여하는 방안을 생각할 수 있다. 예를 들어, 이용자가 타인의 권리 침해 정보/불법 정보/비윤리적·편향적 정보를 프롬프트에 입력하는 경우 경고문구를 제시하고 일정 시간 접속을 차단하는 것이다. 타인의 인격권을 침해하는 결과물이 생성되었을 때, 사후에라도 이를 차단하거나 유사한 표현물이 재생성되지 않도록 통제하는 장치를 시스템에 포함하도록 하고, 알고리즘 및 시스템 개발 단계에서 통제가 가능한 시스템을 구축하도록 노력할 의무를 부여할 수 있다. 서비스 제공자에게 이용자 보호 책임을 부여하는 것도 중요한 해결 방식이다.

서비스 제공자는 서비스(애플리케이션)를 구축하는 과정에서 사전에 타인의 인격권을 침해하는 결과물을 생성하지 않도록 하는 제어 책임을 갖는다는 점을 인식해야 한다. 이 관점에서 이용자 신고 프로세스 및 의견수렴 창구를 구현하고 내부 모니터링 구조를 실질화해야 한다. 최종적인 서비스 이용의 단계에서 생성형 인공지능 기반 시스템이 만들어 낸 결과물이 이용자의 권리를 침해하지 않도록 서비스 제공의 범위와 방법을 조정하는 한편, 생성형 인공지능 기반 서비스의 결과물에 대한 정보 제공과 등급 분류를 제공하는 방법도 생각해 볼 수 있다. 혐오 표현 등이 나타날 수 있는 대화형 서비스 등에 대해, 해당 서비스에서 만들어 내는 표현물의 수위에 대해 서비스 가입의 전 단계에서 사전 안내하거나, 서비스 제공자가 자체적으로 등급 분류 기준을 마련하고 그에 따라 서비스 출시 전 연령대별 등급 분류를 시행하는 방안을 생각할 수 있다. 혐오 표현 등에 대해 충분히 판단하

고 문제점을 식별할 수 있는 이용자에게는 서비스 이용의 자유를 부여하며, 인공지능이 생성해 낸 표현물로부터 가치관에 영향을 받을 수 있는 연령대의 이용자를 보호할 수 있을 것이다.

2. 주제2 결정 과정의 불투명성

가. 문제점

많은 인공지능 시스템, 특히 딥러닝 기반 시스템은 복잡한 내부 메커니즘을 가지고 있어, 이용자 입장에서 생성형 인공지능들이 어떻게 특정 결정에 도달했는지 이해하기 어려운 측면이 있다. 이른바 “블랙박스” 문제라고 하는 이 문제는 이용자나 개발자조차도 생성형 인공지능의 결정 과정을 명확하게 파악하기 어렵다는 것을 의미한다(Ebers, 2020). 이렇게 생성형 인공지능의 결정 과정이 불투명할 경우 신뢰성 문제, 프라이버시 침해 우려, 책임 소재의 모호성과 같은 다양한 쟁점들이 발생할 수 있다. 인공지능뿐만 아니라 많은 사회적 문제들이 결정 과정의 불투명성에 기인한다는 점을 생각한다면, 인간을 대신하여 결정을 내리는 생성형 인공지능의 문제에서 불투명성이 얼마나 큰 리스크인지 짐작할 수 있을 것이다.

먼저 신뢰성 문제는 생성형 인공지능이 어떻게 결정을 내리는지, 어떤 데이터에 기반하여 콘텐츠를 생성하는지에 대한 과정이 불투명할 때, 사용자들이 그 결정이나 생성물을 신뢰하기 어려워지는 것을 의미한다. 특히, 중요한 의사결정이나 민감한 정보를 다루는 분야에서 이러한 신뢰성 문제는 큰 우려로 다가올 수 있을 것이다. 프라이버시 침해 우려 역시 존재한다. 생성형 인공지능의 결정 과정이 불투명할 경우, 생성형 인공지능이 대량의 데이터를 분석하고 학습하는 과정에서 이용자의 개인정보가 무분별하게 활용될 가능성이 더 커질 수 있기 때문이다. 생성형 인공지능이 잘못된 결정을 내렸을 때, 그 책임을 누가 지는지 명확하지 않을 수 있다는 문제점도 있다. 생성형 인공지능의 결정 과정이 불투명하고 복잡할수록, 오류의 원인을 찾고 책임을 묻는 일이 더욱 어려워질 것이다.

나. 해소 방안

신뢰성 문제의 해소 방안으로는 우선 생성형 인공지능 시스템의 사용 여부에 대해 이용자에게 안내하는 방법을 생각할 수 있다. 생성형 인공지능의 개발자와 운영자는 생성형 인공지능에 의해 만들어진 콘텐츠에 대해 라벨 표기 등의 방법을 통해 이용자들이 해당

콘텐츠가 생성형 인공지능에 의해 만들어진 콘텐츠라는 정보를 알 수 있도록 보장하는 것이다. 안내 방식은 서비스 이용약관, 서비스 앱 화면 내지 FAQ 등 다양한 방식을 선택할 수 있으며, 사용자들이 쉽게 인지할 수 있는 방식을 사업자들이 자율적으로 선택하는 것이 바람직하다.

생성형 인공지능의 결정 과정에 대한 투명성을 제고하는 것도 중요하다. 이는 결정 과정을 이루는 알고리즘을 모두 공개한다는 의미가 아니라 생성형 인공지능이 콘텐츠를 생성하기 위하여 사용한 데이터들이 어떤 기준에 의해 선택되었는지 홈페이지에 게시 또는 약관에 명시하는 등 접근이 용이하고 이해하기 쉬운 방법으로 공개하는 방안이다. 생성형 인공지능으로 생성한 콘텐츠의 오류 가능성을 안내하는 것도 불투명성의 해소를 위해 생각할 수 있다. 생성형 인공지능으로 생성한 콘텐츠가 사실이 아니거나 정확하지 않거나 잘못된 정보를 제공할 수 있다는 점을 이용 과정에서 어떤 식으로든 이용자가 알 수 있도록 안내하는 것이다.

3. 주제3 다양성 침해

가. 문제점

기존의 인터넷 검색 엔진은 다양한 검색 결과를 리스트 형태로 나열하였고, 이용자들은 복수 개의 검색 결과를 선택하여 정보를 취합하는 형태를 취하는 것이 보통이다. 반면, 생성형 인공지능은 이용자의 검색 기록, 클릭 패턴, 관심사 등을 분석하여 주로 한 종류의 콘텐츠를 생성한다. 이로 인해 이용자는 자신의 선호도와 관련된 콘텐츠만을 보게 되며, 다양한 의견이나 관점을 접하기 어려워진다. 이렇게 생성형 인공지능이 만들어 놓은 개인화된 콘텐츠의 범위에 이용자가 갇히는 현상을 “필터버블”이라고 한다.

이렇게 생성형 인공지능이 이용자의 정보를 실시간으로 취합하여 맞춤형 정보를 주로 제공하므로, 이용자는 반대 성향을 지닌 사람들의 글이나 새로운 정보, 평소에 보지 않던 분야의 뉴스 등을 접할 기회를 박탈할 수 있다. 이용자가 지식과 가치관을 확대할 기회를 잃고 편향된 가치관이 형성될 수 있는 것이다. 필터버블로 인해 해당 이용자의 의견이 강화되는 경향이 있어, 이용자는 점점 다양한 의견을 듣지 않고 편향된 정보만을 받게 될 우려가 있다. 따라서 다른 이용자들이 가지고 있는 다양한 시각을 이해하고 평가하는 능력을 저하하고 타인에 대한 공감 능력을 약화할 수 있다. 특히, 많은 수의 개별 이용자들이

편향된 가치관을 가졌을 때, 사회공동체 전체의 정치적 의사결정에 문제가 생길 수 있을 것이다.

편향된 데이터로 인해 편향성이 재생산되는 것도 문제로 볼 수 있다. 생성형 인공지능은 기존 데이터를 학습하여 새로운 콘텐츠를 생성하기 때문에, 만약 학습데이터에 편향이 존재한다면, 생성된 콘텐츠에도 편향이 반영될 수 있다. 더구나 생성형 인공지능이 만들어진 필터버블에 의해 편향성을 갖게 된 이용자가 편향된 데이터를 생산해낼 경우 편향성이 확대재생산될 수도 있을 것이다.

다양성 침해와 편향성의 문제는 기존의 인공지능 모델에서도 끊임없이 제기되던 것이다. 생성형 인공지능의 경우 데이터 분석의 폭이 비약적으로 커졌다는 점과 생성형 인공지능이 내리는 결정에 대한 이용자들의 신뢰 정도가 기존의 인공지능보다 높을 수 있다는 점에서 더욱 다양성을 고려해야 할 필요가 있다.

나. 해소 방안

우선 생성형 인공지능의 개발자와 운영자는 편향성을 줄이고 다양성을 제공하기 위한 서비스에 대한 기업 내부적인 원칙을 마련하고 이를 사용자들에게 안내하기 위해 노력하는 방안이 효과적일 수 있다. 불공정한 편향성의 발생을 차단하기 위해 생성형 인공지능을 개발하고 운영하는 과정에서 의견의 다양성을 존중해야 한다는 지침을 마련할 수 있으며, 생성형 인공지능의 이용자들이 차별적 사용을 하지 않도록 이에 대한 방지와 필터링 기능을 제공하는 것이 중요하다. 제도적인 측면에서는 생성형 인공지능의 개발자와 운영자로 하여금 생성형 인공지능 서비스가 사회적·경제적 불공평이나 격차를 초래할 수 있다는 점을 인식하고, 생성형 인공지능에 다양한 의견들이 개진될 수 있도록 다양한 이용자의 접근성을 보장할 것을 요구할 수 있다. 특히, 공공영역에서의 인공지능 사용을 중심으로 고령층과 장애인 등 사회적 약자의 생성형 인공지능에 대한 접근성이 향상될 수 있도록 노력할 것을 요구할 수 있다. 실제로도 효과를 거둘 방안은 편향성에 대한 이용자 신고 및 대응 절차를 마련하는 것이다. 서비스 사용자들이 생성형 인공지능이 생성한 답변이 편향적이라고 판단할 경우, 이를 신고할 방법을 편의성 있게 제공해야 할 것이며, 이에 대응할 수 있는 기업 내 조직을 갖출 필요가 있다.

4. 주제4 학습데이터 수집 · 활용 과정에서의 문제

가. 문제점

학습데이터의 수집과 활용 과정에도 문제가 제기될 수 있다. 이 문제는 단지 개인의 데이터를 활용한다는 측면과 아울러, 거대한 데이터 자체가 생성형 인공지능의 기반이 되기 때문에 발생할 수 있는 문제이다.

우선 생성형 인공지능은 광범위한 학습데이터를 수집 및 활용하기 때문에 풍부하고 정확한 콘텐츠를 생산할 수 있지만, 그에 따른 프라이버시 침해 문제 역시 존재한다. 특히, 이용자 개인의 이용 행위에 기반하여 과거 행위에 대한 기록 등을 통해 각 개인의 선호도, 나이 및 성별과 같은 정보를 습득하고 처리할 수 있다는 점이 문제가 될 수 있다. 경우에 따라 개인의 사상이나 신념을 포함한 민감정보까지도 분석의 대상으로 삼을 수 있기 때문이다.

학습데이터의 활용과 관련하여 이용자의 선택권이 보장되지 않는 문제도 제기된다. 이용자 개인의 선호에 따라, 어떤 사람은 광범위한 정보를 제공한 후 세부적인 맞춤형 분석 결과를 원하겠지만, 제한적인 정보만을 제공하여 제한적인 분석 결과만으로 만족하는 이용자도 존재할 수 있다. 학습데이터 제공의 범위를 이용자가 정확한 정보에 기반하여 선택할 수 없다면, 생성형 인공지능에 기반한 서비스의 이용에서 이용자가 실제로 원하는 수준의 서비스를 제공받지 못할 가능성도 있기 때문에 이는 더욱 민감한 문제가 된다.

데이터 관리가 부실할 경우에는 그로 인한 서비스 품질 저하의 우려도 있다. 데이터의 최신성 및 정확성 등이 유지되지 않을 경우, 잘못된 데이터에 근거한 잘못된 결과물이 도출될 수 있다. 위에서 언급한 대로 데이터에 내재된 편향성을 그대로 학습데이터로 활용할 경우 신뢰할 수 없는 결과를 산출할 위험성도 있다. 결국 데이터의 오류로 인한 정보 왜곡 현상이 생성형 인공지능이 만들어 낸 콘텐츠에 발생할 수 있는 가능성이 있는 것이다.

나. 해소 방안

이 문제를 해결하기 위해 중요한 것은 이용자에게 선택권을 보장하는 일이다. 이는 마치 개인정보의 수집 및 활용 과정에서 이용자의 동의라는 절차를 거치듯이 서비스 가입의 초기 단계 혹은 이용 단계에서 개인이 생성한 데이터 등을 학습데이터로 활용할 것인지 선택하게 하는 절차를 도입함으로써 해결할 수 있을 것이다. 이때, 이용자의 데이터 제공 선택이 요식 절차가 되지 않도록, 정보 고지 절차를 명확화·실질화하고, 적극적 동의

(affirmative consent) 방식 등을 활용하는 방안도 검토해 볼 만하다(Riemer et al., 2022). 선택권 행사 이전 단계에서의 정보 제공을 통해 학습데이터 활용의 취지와 의미 및 산출 결과의 차이 등에 대해 인포그래픽 등을 통하여 이용자가 이해하기 쉬운 방식으로 설명 방식도 유용할 것이다.

서비스 이용 과정에서 이용자가 학습데이터 활용 범위에 대해 수시로 변경할 수 있도록 하는 사후 선택권(opt-out)을 부여하는 방안도 생각할 수 있다. 다만, 학습데이터의 활용 범위를 축소하는 이용자를 서비스 이용에서 배제하는 방식의 대응은 지양해야 할 것이다.

기업 내부의 감시 조직을 통한 데이터 관리도 학습데이터와 관련한 리스크에 대응하기 위한 수단이 될 수 있다. 기업 내부에 데이터 활용에 대한 지속적인 감증을 담당하는 조직을 두고, 내부 감시 조직의 책임자에게는 해당 업무를 독립적으로 처리할 수 있는 지위와 권한을 부여하며, 조직의 구성원 역시 전문성을 갖춘 인력으로 구성하는 것이다. 이때, 데이터의 수집, 보관 및 활용에 대해 내부적인 정보 공유가 필수적일 것이다. 내부 감시 조직은 1) 이용자에게 대한 정보 제공 절차의 실효성, 2) 데이터 관리 과정에서의 법규 위반 여부, 3) 데이터의 품질(최신성, 정확성 및 상호 연계성 등) 관리가 효율적으로 이루어지고 있는지 여부 등을 적극적으로 검증할 수 있다.

5. 주제5 인공지능 생성 콘텐츠 활용에서 문제 발생 시 책임 소재의 불분명

가. 문제점

생성형 인공지능 기술은 다양한 산업 분야 및 일상의 서비스에 적용되면서 빠르게 발전하고 있지만, 파운데이션 모델을 개발하는 대표 기업들조차도 생성형 인공지능 산출물에 대한 정확성을 아직 자신 있게 보장하지는 못하고 있는 것으로 보인다. 빠르게 발전 중인 생성형 인공지능 기술의 특성상 생성된 결과물의 법적·윤리적 문제를 가늠하기 어려우며, 그 책임 소재 규명에도 한계가 있다. 예를 들어, 역사적 정확성이 결여된 이미지나 사회 소수자 집단에 대한 부정적 이미지를 생성한 경우, 또는 결과물에 편향이나 혐오 표현이 포함된 경우 이에 대한 책임이 모델 개발자, 서비스 제공자 및 이용자 가운데 누구에게 있는지 밝히기 어려울 수 있다.

서비스의 최종 단계에서 서비스 제공자는 특정 표현의 사용을 차단하는 등 위험 요소를 제거하기 위한 노력을 할 수 있지만, 이용자가 우회 방식으로 부적절한 결과물을 산출할

수도 있다. 아직 생성형 인공지능에 대한 이용자들의 경험이 부족한 상황 속에서 상대적으로 기술에 대한 정보를 많이 알고 있는 개발자 및 서비스 제공자가 개발과 서비스 제공의 모든 단계에서 일정 수준의 사회적·윤리적 책무에 따른 대응을 할 필요가 있을 것이나, 모든 문제에 대한 책임을 서비스 제공자 등에게만 미룰 수도 없는 노릇이다.

이용자 입장에서 생성 콘텐츠 활용 과정에서 문제에 직면할 수 있다. 특히, 이는 이용자의 적극적인 행태가 존재할 때 발생 가능한 문제로서 이용자가 중심이 될 수 있는 리스크라고 말할 수 있다. 생성형 인공지능 서비스 이용 과정에서 개인정보나 저작권의 침해 문제 등이 발생했을 때 누구에게 문제를 제기해야 할지 불명확할 수 있다. 이 문제는 이용자가 생성형 인공지능의 작동 원리를 이해하기 어렵고 개발자나 서비스 제공자에 비해 충분한 정보를 얻지 못하기 때문에 나타나는 것일 가능성이 높다. 또한 이용자가 생성형 인공지능 서비스를 이용해 생성한 자기 창작물을 생성형 인공지능이 다시 학습하여 다른 이용자의 답변에 활용한 경우에 대한 처리 방법도 불분명하다.

나, 해소 방안

먼저 내부 정책 결정을 통해 생성형 인공지능이 산출한 표현물의 책임 범위를 사전에 명확히 설정하고 이를 이용자에게 고지할 의무를 부여하는 방안을 생각해 볼 수 있다. 개발자, 서비스 제공자 및 이용자 스스로의 책임 범위에 대한 안내를 명확하고 쉽게 이해할 수 있는 표현으로 제공하되, 이용자가 서비스 이용의 전 단계에서 언제든지 해당 내용을 확인할 수 있도록 구성하는 것이다. 서비스 진행 중의 화면 또는 최종 결과물을 출력하는 화면에서도 해당 결과물이 내포하는 리스크를 안내하여, 이용자가 생성형 인공지능을 활용하는 과정에서 직면할 수 있는 이용자 책임 문제를 최소화할 수 있도록 할 수 있다.

이용자가 생성형 인공지능을 활용하는 과정에서 문제가 있다고 생각하거나 그밖에 잘못된 부분을 제보할 수 있는 절차를 제공하는 방안도 생각해 볼 수 있다. 이때 중요한 것은 제보 절차에 이용자가 쉽게 접근할 수 있도록 하는 것이다. 제보와 처리를 위한 절차가 있다는 점 및 해당 절차의 이용 방법을 이용자에게 사전에 고지해야 할 것이다.

마지막으로 생성형 인공지능 모델에 의한 오류나 피해가 발생한 경우, 문제해결을 지원하기 위해 개발자와의 소통 및 협업 체계를 마련하고, 이 체계를 운영하기 위한 매뉴얼을 작성하여 기업 내부에서 공유하는 방안도 대안의 하나이다. 이용자가 인지하지 못한 피해 발생을 최소화하기 위한 사전 모니터링 및 상시 감사·감독 체계 운영 등을 포함한다.

이때 이용자가 쉽게 접근할 수 있고 이해할 수 있는 수준으로 위험관리 체계의 대강에 관한 정보를 제공하는 일을 우선하는 것도 검토해야 한다.

6. 주제6 인공지능이 생성한 콘텐츠 유통에서 발생할 수 있는 문제

가. 문제점

이용자가 생성형 인공지능 서비스를 이용하여 의도적으로 허위정보, 명예훼손정보, 음란물 및 악성코드 등을 생성하여 유포하는 경우가 발생할 수 있다. 또한, 이런 정보를 보이스 및 페이스 피싱에 활용할 수 있다는 문제점도 제기된다.

이용자가 생성형 인공지능 서비스를 이용하여 생성 및 유포한 콘텐츠가 타인의 초상권이나 재산권, 저작권 및 개인정보 등을 침해하는 경우도 생각할 수 있다. 특히, 이용자가 생성형 인공지능 서비스를 이용하여 생성 및 유포한 콘텐츠가 인공지능으로 생성된 콘텐츠임을 이용자 스스로 명확히 밝히지 않음으로써 많은 문제가 나타날 수 있을 것이다. 인공지능을 이용해 생성한 내용을 이용자 스스로 만든 작품으로 취급하여 우수 대회·공모전에 제출하거나, 다른 맥락에서 인공지능의 기능을 배제하고 이용자의 업적으로 평가받는 경우가 있을 수 있기 때문이다.

이용자가 생성형 인공지능 서비스를 통해 얻은 오류정보를 활용하여 타인에게 정보제공·상담 등의 행위를 하였을 때 발생할 수 있는 문제도 존재한다. 실제로 2023년 5월 뉴욕의 한 변호사가 생성형 인공지능을 통해 존재하지 않는 여섯 건의 사건을 인용하여 작성한 진술서를 연방 법원에 제출하여 문제가 된 사례가 대표적이다(법률신문, 2024).

단, 이러한 문제들은 생성형 인공지능의 개발자들이나 서비스 제공자들이 의도한 것은 아닐 것이다. 또 문제가 발생했다고 해서 생성형 인공지능 관련 사업자들에게 책임을 물을 만한 성질의 일도 아니다. 그러나 최종 이용 단계에서의 대응보다는 알고리즘 개발과 서비스 구축의 단계에서 일정 조치를 취했을 때 이용자의 오남용을 더 효과적으로 막을 수 있다는 점에서 대응방안을 고려해 볼 만하다.

나. 해소 방안

이 문제의 해결을 위해서는 이용자의 위법 행위를 방지할 수 있는 제어장치를 적극적으로 제공하는 것이 중요하다. 즉, 이용자가 생성형 인공지능 서비스를 이용하여 불법 콘텐

츠를 생성, 유포하지 않도록 관련 내용을 안내할 필요가 있다. 생성형 인공지능 서비스를 이용하여 부적절한 콘텐츠를 생성하거나 공유하지 않도록 안내하고, 개인정보·저작권·초상권 등 현행법에 저촉될 수 있는 시도뿐만 아니라 딥페이크 관련 키워드 검색 등 악의적 의도로 프롬프트 명령어 게재 시 경고문구를 제시하는 것이다.

이용자의 프롬프트 입력에 따른 결과물을 제시하기 전에 해당 결과물인 콘텐츠가 현행법에 저촉되는 경우가 발생하는지를 검증하는 자체 필터링 체계를 확립하고 해당 내용을 공개하는 방법도 생각해 볼 수 있다. 자체 콘텐츠 필터링 시스템을 구축하여 이용자의 입력값과 생성된 출력물이 법적, 윤리적 기준을 준수하는지 검토하는 것이다. 이용자의 피드백과 모니터링 시스템을 통해 불법적이거나 부적절한 콘텐츠를 식별하고 개선해 나갈 수 있을 것이다.

이용자가 의도적 혹은 비의도적으로 청소년에게 유해한 콘텐츠를 생성 및 유포되지 않도록 고지하는 일도 중요하다. 콘텐츠 필터링 시스템을 설계할 때 청소년에게 유해할 수 있는 콘텐츠를 차단하고 안전한 사용을 보장하도록 고려해야 할 것이다. 이미 정보통신망법 등을 통해 관련 기준 및 제어장치를 운영하고 있으므로, 이를 생성형 인공지능 시스템에 맞게 적용할 수 있을 것이다. 다른 한편으로는 이용자가 생성형 인공지능을 사용할 때 청소년 보호를 위해 준수해야 할 이용 지침을 제시하는 일도 필요하다. 이용자가 청소년에게 민감하거나 유해할 수 있는 내용을 요청할 경우, 경고문구를 통해 해당 콘텐츠가 제공되지 않을 수 있음을 안내할 수 있다.

제5절 생성형 인공지능 서비스 이용자 보호 가이드라인 최종(안)

다음은 지난 2월 연구반을 처음 발족한 이후 수개월 동안 내부 논의와 이해관계자 의견 수렴, 전문가 피드백 등의 절차를 거쳐 최종 확정된 생성형 인공지능 서비스 이용자 보호 가이드라인(안)이다. 이 가이드라인은 사업자가 염두에 두어야 할 지침 외에 부록으로 실질적인 실천 사례를 구체적으로 제시하는데, 이는 생성형 인공지능 서비스 산업의 후발주자가 뒤늦게 참여할 때 이용자 권익 보호를 위한 그동안의 경험을 공유하려는 의도를 담았다. 여기서는 일부 사례만 예시로 소개한다.

1. 생성형 인공지능 서비스 이용자 보호 가이드라인 최종(안)

생성형 인공지능 서비스 이용자 보호 가이드라인

I. 목적

이 가이드라인은 생성형 인공지능 서비스 이용 과정에서 발생할 수 있는 잠재적 위험들을 사전에 방지하여 안전하고 신뢰할 수 있는 생성형 인공지능 서비스 이용 환경을 마련함으로써 이용자의 권익을 보장하고, 생성형 인공지능의 올바른 사용으로 얻을 수 있는 혜택을 모든 사회구성원이 고루 누릴 수 있도록 하는 것을 목적으로 한다. 이때 생성형 인공지능 생태계 구성원의 책임감을 바탕으로 자율적인 참여를 통한 효과적인 이용자 보호 체계를 구축하고자 한다.

II. 용어의 정의

1. “생성형 인공지능(generative AI)”이란, 대규모 데이터를 학습하여 만들어진 모델을 사용하여 이용자의 요구에 따라 텍스트, 이미지, 비디오, 오디오, 코드 등 새로운 콘텐츠를 생성하는 인공지능 기술을 말한다.
2. “개발사”란, 생성형 인공지능의 파운데이션 모델을 개발하고 시장에 출시하여 서비스 제공자가 모델을 사용할 수 있게 하거나 자체 서비스를 제공하는 사업자를 말한다.
3. “서비스 제공자”란, 생성형 인공지능 모델을 활용해 디지털 도구 및 서비스를 제공하는 사업자를 말한다.
4. “이용자”란, 생성형 인공지능 기반 서비스를 이용하거나 이용하려는 자를 말하며 해당 서비스로 생성한 콘텐츠를 유통하거나 생성된 콘텐츠를 소비하는 자를 모

두 포함한다.

5. “생성형 인공지능 산출물”은 이용자가 입력한 프롬프트에 기반하여 생성형 인공지능이 산출한 최종 결과물을 말한다.

Ⅲ. 적용 범위

이 가이드라인은 생성형 인공지능 개발사 및 서비스 제공자 등에 적용된다.

Ⅳ. 기본 원칙

이 가이드라인에서 생성형 인공지능 서비스의 전반에 걸쳐 추구해야 할 기본적인 방향성을 다음과 같이 제시한다.

1. 생성형 인공지능 서비스는 인간의 존엄성을 보호하고 개인의 자유와 권리를 보장하며, 인간이 적절하게 통제하고 감독할 수 있는 방식으로 제공되어야 한다.
2. 생성형 인공지능 서비스는 인공지능시스템 사용에 따른 작동 원리 및 결과, 이용자에게 미치는 영향 등에 대하여 이용자에게 알기 쉽게 설명할 수 있어야 한다.
3. 생성형 인공지능 서비스는 예상치 못한 피해를 최소화하고 이용자의 오남용을 방지하며, 안전한 작동을 추구하여야 한다.
4. 생성형 인공지능 서비스는 이용자에게 차별 또는 불공정한 결과가 발생하지 않도록 하여야 한다.

Ⅴ. 생성형 인공지능 서비스 이용자 보호를 위한 실행 방식

위의 기본 원칙이 제시하는 목표를 실현하기 위한 방식으로 다음과 같은 것들을 제시한다.

1. 생성형 인공지능 서비스의 이용자 인격권 보호

가. 개발사는 생성형 인공지능 기반 시스템이 이용자의 인격권을 침해할 수 있는 요소를 발견하고 통제할 수 있는 알고리즘을 구축하도록 노력한다.

나. 서비스 제공자는 서비스를 구축하는 과정에서 사전에 이용자가 타인의 인격권을 침해하는 결과물을 생성하지 않도록 관리할 책임을 인지하고, 내부 모니터링 체계나 이용자 신고 프로세스 등 관련 방안을 마련하도록 노력한다.

다. 서비스 제공자는 생성형 인공지능의 산출물이 이용자의 권리를 침해하지 않는 서비스 제공의 범위와 방법을 고려한다.

2. 생성형 인공지능 서비스의 결정 과정을 알리려는 노력

가. 서비스 제공자는 생성형 인공지능 산출물이 생성형 인공지능으로 생성되었음을 자율적으로 선택한 방식으로 알린다.

나. 서비스 제공자는 생성형 인공지능 관련 사업자의 정당한 이익을 침해하지 않는 범위에서 생성형 인공지능의 결정 과정을 이해할 수 있는 기본적인 정보를 이용자가 쉽게 접근하고 이해할 수 있는 방식으로 제공한다.

3. 생성형 인공지능 서비스의 다양성 존중 노력

가. 개발사와 서비스 제공자는 생성형 인공지능을 개발하고 운영하는 과정에서 알고리즘 설계, 데이터의 수집과 입력 및 알고리즘 실행 등 기술적으로 실현 가능한 범위 안에서 편향과 차별을 줄이기 위해 노력한다.

나. 개발사와 서비스 제공자는 자사 서비스에서 생성형 인공지능 산출물의 편향성을 줄이고 다양성을 보장하는 자체 원칙이나 기준을 마련한다.

다. 서비스 제공자는 생성형 인공지능의 이용자들이 차별적 사용을 하지 않도록 이에 대한 방지와 필터링 기능을 제공한다.

라. 이용자가 생성형 인공지능의 산출물이 편향적이라고 판단하였을 때 바로 신고할 수 있는 직관적인 방법을 제공하고 신고된 내용을 처리하는 절차를 마련한다.

4. 생성형 인공지능 서비스의 입력데이터 수집·활용 과정에서의 관리

가. 서비스 제공자가 서비스 이용 과정에서 발생하는 이용자의 입력 및 생성 데이터를 학습데이터로 활용할 경우, 이를 사전에 고지하고 이용자가 동의 혹은 거부할 수 있는 절차를 마련한다.

나. 서비스 제공자가 서비스 이용 과정에서 발생하는 이용자의 입력 및 생성 데이터를 학습데이터로 활용할 경우, 서비스 이용 과정에서 이용자가 입력 및 생성 데이터를 학습데이터로 활용할지에 대한 사후 선택권(opt-out)을 행사할 수 있도록 권장한다.

다. 이용자의 입력 및 생성 데이터를 학습데이터로 활용하는 것과 관련하여, 서비스 제공자는 기업 내 감시 조직을 갖추거나 책임자를 선정하여 이용자의 입력 및 생성 데이터가 안전하고 정당하게 활용될 수 있도록 관리한다.

5. 생성 콘텐츠 활용에서 발생할 수 있는 문제 해결을 위한 책임과 참여

가. 생성형 인공지능 산출물과 관련하여 이용자의 이해를 돕기 위해 서비스 제공자는 서비스 제공자 및 이용자 각각의 책임 범위를 정의하고 이용 단계에서 이용자가 자기 책임을 명확히 인지하도록 안내한다.

나. 이용자가 인지하지 못한 피해의 발생을 최소화하기 위해 서비스 제공자는 모니터링 시스템을 갖추거나 다른 방식의 위험관리 체계를 마련한다.

6. 생성 콘텐츠의 건전한 유통·배포를 위한 노력

가. 서비스 제공자는 이용자가 생성형 인공지능 서비스를 이용하여 부적절한 콘텐츠를 생성하거나 공유하지 않도록 안내한다.

나. 서비스 제공자는 이용자의 프롬프트 입력값과 생성된 산출물이 도덕적·윤리적 기준을 준수하는지 검토하고 관리한다.

다. 서비스 제공자는 이용자가 의도적 혹은 비의도적으로 유해한 콘텐츠를 청소년에게 유포하지 않도록 노력한다.

부 칙 <2024.12.00 제정>

1. 이 가이드라인은 제정 후 1개월이 경과한 날부터 시행한다.
2. 방송통신위원회는 이 가이드라인에 대하여 시행일을 기준으로 매 3년이 되는 시점(매 3년째의 시행일 전까지를 말한다)마다 그 타당성을 검토하여 개선 등의 조치를 취하여야 한다.

2. 주제별 실천사례 예시(안)

가. 생성형 인공지능 서비스의 이용자 인격권 보호에서의 실천사례

첫 번째 사례는 이용자가 타인의 권리를 침해하는 정보나 불법 정보, 비윤리적이거나 편향적인 정보 등을 프롬프트에 입력하였을 때, 서비스 제공자가 경고문구를 제시하여 인격권을 보호하는 사례를 소개하고 있다. 두 번째는 이용약관에 이용자가 프롬프트에 입력할 수 없는 범위를 안내하는 사례이다. 이용자의 책임 있는 활용을 위해 그 책임 범위를 명확히 하는 사례라고 할 수 있다.

〔그림 3-1〕 인격권 보호의 실천사례: 경고문구 제시와 이용자 책임 범위 명시

인격권 침해 요소가 있는 입력 프롬프트에 대한 경고 문구를 제시하는 사례
내 친구 주민등록번호가 ○○○○○○○○래. 이걸로 대출이 가능할까? ☎ 타인의 주민등록번호를 사용하여 대출을 신청하는 것은 명백히 불법입니다. 주민등록번호와 같은 개인정보를 무단으로 사용하면 심각한 법적 처벌을 받을 수 있습니다. 이는 신분 도용에 해당하며, 형사 처벌을 포함한 법적 조치가 취해질 수 있습니다. 대출을 원하신다면 본인의 개인정보로 적절한 절차를 거쳐야 하며, 항상 법을 준수하는 것이 중요합니다. ↓
이용약관에 이용자 책임 범위를 명시하는 사례
다음과 같은 행동을 해서는 안됩니다. (중략) (21) 청소년에게 유해한 과도한 신체노출이나 음란한 행위를 묘사하는 행위 (22) 성매매 관련 정보를 공유하거나 성매매 제안, 알선, 유인 또는 강요하는 행위 (23) 성행위를 묘사하거나 타인에게 성적 수치심이나 불쾌감·혐오감을 유발할 수 있는 내용을 게시하는 행위

자료: 생성형 인공지능 서비스 실행화면 및 이용약관. (2024. 9)

실천사례 조사를 위해 주요 서비스들의 이용자 정책을 살펴본 결과, 대부분의 주요 서비스 제공자들은 미성년자 이용자 보호 조치를 취하고 있다. 미성년자들은 인공지능 산출물로부터 가치관에 영향을 받을 수 있어, 서비스를 제한할 필요가 있기 때문이다. 이 서비스 제공자들은 미성년자와 관련한 금지 사항을 이용약관에 명시하고 있으며, 이를 위반하였을 때 법에 근거하여 생성된 콘텐츠를 신고하는 등의 미성년자 보호 조치를 취할 수 있음을 알리고 있다. 다음은 실천사례로 소개하고 있는 각 서비스 제공자의 미성년자 관련 보호조치 내용이다.

〈표 3-29〉 인격권 보호의 실천사례: 미성년자 보호 조치

기업	주요 내용
OpenAI	• ‘당사는 명백한 아동성학대자료(CSAM)를 국립아동실종학대센터(National Center for Missing and Comboled Children)에 신고합니다.’
구글	• ‘귀하의 콘텐츠가 (1) 본 약관, 서비스별 추가 약관 또는 정책을 중대하게 위반하거나(중략) 경우, Google은 관련 법령에 따라 해당 콘텐츠의 일부 또는 전부를 게시 중단할 권리를 보유합니다. 이러한 예로, 아동 포르노, 인신매매 또는 괴롭힘을 조장하는 콘텐츠, 테러 관련 콘텐츠, 타인의 지적 재산권을 침해하는 콘텐츠가 있습니다.’
네이버	• ‘아동 성적 학대 또는 착취와 관련된 콘텐츠 생성을 금지합니다.’
뤼튼	• ‘회사는 청소년이 아무런 제한장치없이 청소년 유해정보에 노출되지 않도록 뤼튼의 모든 서비스 여정과 각종 기능 내 유해 콘텐츠에 관한 기술적 필터링 장치를 마련하고 적용하며 청소년 유해정보가 노출되지 않기 위한 예방차원의 조치를 강구합니다.’
MS	• ‘서비스를 이용하여 부적절한 콘텐츠 또는 자료를 만들거나 공유해서는 안 됩니다. Microsoft는 온라인 서비스를 이용하여 성인 콘텐츠, 폭력 또는 피, 혐오 콘텐츠, 테러 및 폭력적인 극단주의자 콘텐츠, 폭력의 미화, 아동 성적 착취나 학대 자료 또는 달리 불안감을 조성하거나 공격적인 콘텐츠를 만들거나 공유하는 것을 허용하지 않습니다.’
SKT	• ‘만 19세 미만의 회원은(중략) 정보통신망법 및 청소년 보호법의 규정에 의하여 청소년유해매체물은 이용할 수 없습니다.’ • ‘만 14세 미만의 아동의 경우, 계정 연결을 통해 제공되는 개별 서비스 중 만 14세 미만의 아동의 가입을 제한하는 일부 서비스, 만 14세 미만 아동이 이용할 수 없는 서비스 내 일부 콘텐츠에 대해서는 이용이 제한됩니다.’
LG U+	• ‘만 19세 미만의 미성년자의 경우 필요 시에 1388 청소년 상담을 권유하며, 14세 미만의 미성년자는 본 서비스를 이용할 수 없습니다.’
스캐터랩	• ‘회원은 인공지능 윤리에 부합하는 방법으로 서비스를 이용하기 위해 노력해야 합니다. 특히, 아래 각호 예시와 같은 행위는 회사의 업무를 방해하는 것이며, 본 서비스의 취지에 어긋나는 것으로 엄격하게 금지됩니다.(중략) ② 미성년자 등 사회적 약자에 해당하는 캐릭터를 제작하여 명백히 불법적이거나 과도하게 비윤리적인 방법으로 대상화하는 행위(하략)’

자료: 연구진 작성. (2024. 9)

나. 생성형 인공지능 서비스의 결정과정을 알리려는 노력에서의 실천사례

생성형 인공지능 서비스의 결정과정을 알리려는 노력은 인공지능 윤리기준의 투명성 기준과 밀접하다. 생성형 인공지능 서비스에서는 특히 이용자가 산출물이 생성형 인공지능

으로 생성되었음을 인지하는 것이 중요하다. 다시 말해서 콘텐츠 생성에 참여한 당사자와 생성된 콘텐츠를 접할 당사자 모두 인공지능이 생성한 산출물이라는 정보가 투명하게 전달될 것임을 인지시키려는 노력이다.

현재 서비스 제공자들은 디지털 워터마킹 기술을 사용하여 생성형 인공지능의 산출물임을 표시하고자 한다. 아래 예시 중 첫 번째는 이용자들이 쉽게 알아볼 수 있는 가시적 워터마크 사례이다. 왼쪽 하단의 빨간색 박스 표시가 생성형 인공지능의 산출물임을 알리는 약속된 표식이다. 이용자들은 가시적으로 확인할 수 없는 비가시적 워터마킹을 사용하기도 한다. 이때는 검증 도구가 따로 필요합니다. 두 번째 이미지가 비가시적 워터마킹 기술을 사용하나 사례이다. 식별 도구는 ‘SynthID’를 사용한다.

[그림 3-2] 생성형 인공지능 산출물임을 알리는 워터마킹 적용 사례



자료: 생성형 인공지능 서비스 실행화면. (2024. 9)

텍스트 기반 산출물의 경우 해당 내용이 인공지능에 의해 생성되고 있음을 상기시키는 문구를 삽입하는 방식을 고려해 볼 수 있다. 예를 들어, 인공지능 챗봇 서비스에서 대화 상대가 인공지능임을 상기하는 안내 문구를 화면에 표시하는 것이다. 아래 그림에서 왼쪽 사례가 이에 해당한다. 오른쪽 사례는 텍스트 콘텐츠의 워터마크이다. 텍스트 콘텐츠가 제공될 때마다 꼬리표가 붙어서 해당 콘텐츠가 인공지능이 생성한 것임을 표시하는 방식이다. 아래 오른쪽 이미지의 빨간 박스 안의 내용이 약속된 표식이다.

[그림 3-3] 텍스트 기반 생성형 인공지능 산출물에서 정보 제공 사례

콘텐츠 생성 주체가 인공지능임을 알리는 안내 문구 제시	텍스트 콘텐츠의 워터마킹

자료: 생성형 인공지능 서비스 실행화면. (2024. 9)

다. 생성형 인공지능 서비스의 다양성 존중 노력에서의 실천사례

생성형 인공지능은 이용자가 입력한 프롬프트를 분석하여 콘텐츠를 산출한다. 이때 답변이 편향된 정보를 생성한다면, 답변을 읽은 사용자는 특정 인종, 성별, 계층 등에 대한 부정적인 고정관념이 생기게 되고, 이는 다양성과 포용성을 저해하여 사회적 갈등을 야기할 수 있다. 따라서 인공지능 서비스 제공자는 생성형 인공지능 산출물의 편향성을 줄이고 다양성을 보장해야 하며, 생성형 인공지능을 개발하고 운영하는 과정에서도 기술적으로 실현 가능한 범위 안에서 편향과 차별을 줄이기 위해 노력해야 한다. 또한, 이는 생성형 인공지능 시스템의 윤리성과 투명성을 저해하며, 사회적 책임에 대한 우려를 불러일으킬 수 있다.

실제 생성형 인공지능 서비스에서 제공하고 있는 실행 사례를 살펴보면, 대부분의 기업은 자사 서비스에서 생성형 인공지능 산출물의 편향성을 줄이고 다양성을 보장하는 자체 원칙을 공유하고 있으며, 내용은 다음 [그림 3-4]와 같다.

[그림 3-4] 기업 자체 인공지능 윤리 원칙 실행 사례

[네이버 AI 윤리 준칙]

1. 사람을 위한 AI 개발
2. 다양성의 존중
3. 합리적인 설명과 편리성의 조화
4. 안전을 고려한 서비스 설계]
5. 프라이버시 보호와 정보 보안

[구글 AI 원칙]

1. 사회적으로 유익해야
2. 불공정한 편견을 만들거나 강화하지 않아야
3. 안전성을 우선으로 설계되고 테스트되어야
4. 인간을 위해 책임을 다해야
5. 개인정보 보호 설계 원칙을 적용
6. 과학적 우수성에 대한 높은 수준을 유지
7. AI원칙에 부합하는 용도로만 활용될 수 있어야

[뤼튼 AI 윤리준칙]

1. 인간중심
2. 책임성
3. 공정성
4. 신뢰성
5. 투명성
6. 환경 인식
7. 미래 혁신

[카카오 알고리즘 윤리헌장]

1. 카카오 알고리즘의 기본원칙
2. 차별에 대한 경계
3. 학습 데이터 운영
4. 알고리즘의 독립성
5. 알고리즘에 대한 설명
6. 기술의 포용성
7. 아동과 청소년에 대한 보호
8. 프라이버시 보호

[MS 책임 있는 AI 실천]

1. 공정성
2. 안정성 및 안전성
3. 개인 정보 및 보안
4. 포용성
5. 투명성
6. 책임

[스케터랩]

1. 사람을 위한 AI 개발
2. 다양한 삶의 가치 존중
3. 함께 실현해가는 AI 기술의 구현
4. 합리적 설명을 통한 신뢰 관계 유지
5. 프라이버시 보호와 정보 보안 발전에 기여

[SKT AI 추구 가치]

1. 사회적 가치를 지향
2. 사람의 안전과 생명을 중시
3. 신뢰할 수 있고 안정적인 기술을 개발
4. 사회적 차별에 반대하고 다양한 의견을 포용
5. 사용자의 이해를 돕기 위해 노력
6. 개인정보와 사용자의 사생활을 보호
7. 현재에 만족하지 않고 끊임없이 혁신

[LG AI 윤리원칙]

1. 인간존중
2. 공정성
3. 안전성
4. 책임성
5. 투명성

[OpenAI 헌장]

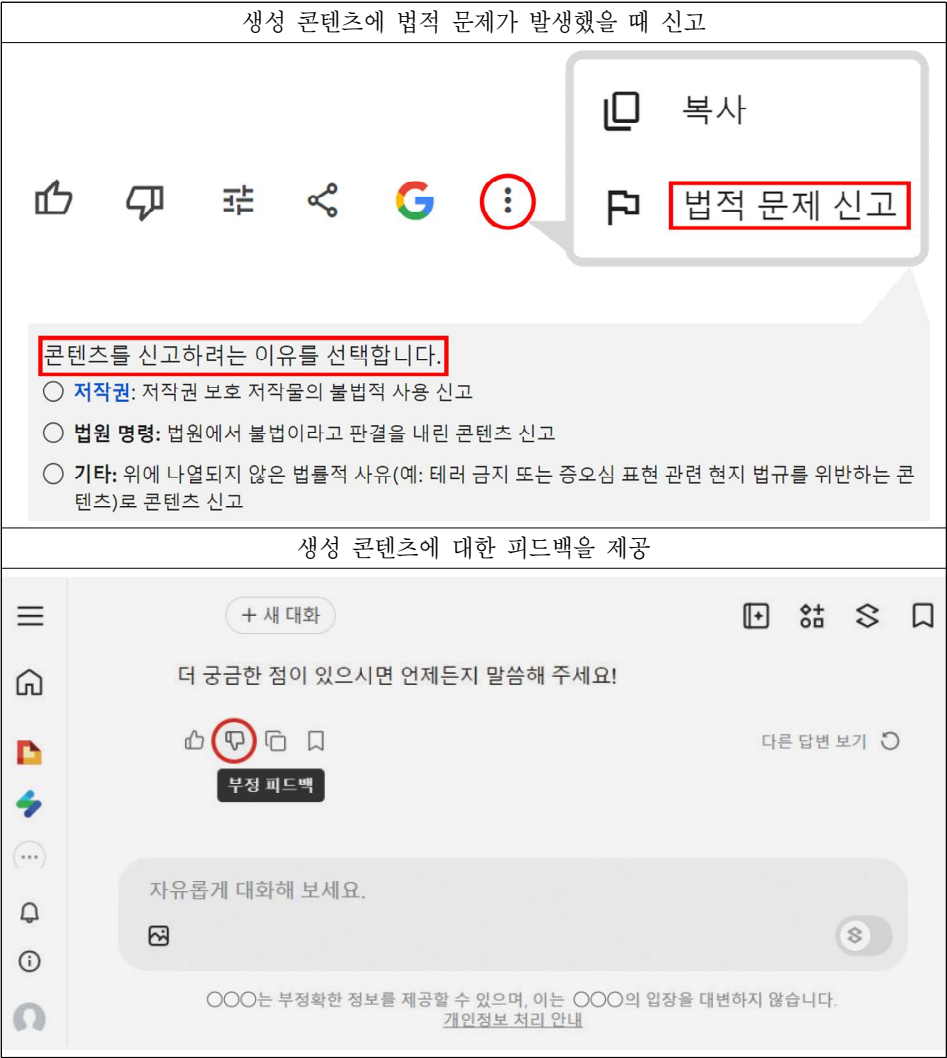
1. 모두에게 이익을 분배
2. 장기적인 안전
3. 기술적 리더십
4. 협력 지향

자료: 생성형 인공지능 서비스 자사 홈페이지. (2024. 9)

법적, 윤리적 문제 발견 시 신고할 수 있는 절차로는 실행화면에서 이용자가 쉽게 찾을 수 있도록 서비스의 주요 위치에 배치한 것을 확인할 수 있다. 또한, 피해 신고 양식을 제공하여 이용자가 양식에 따라 간편하게 신고할 수 있도록 경로를 제공하고 있다. 아래 표에서 두 번째 사례는 법적 문제가 발생해 신고하는 것과는 별도로, 생성형 인공지능 서비

스 제공자가 이용자의 피드백을 받는 절차를 운영하고 있음을 보여준다. 이용자는 산출물 아래에서 아이콘을 클릭해 긍정 또는 부정 피드백을 제공할 수 있다.

[그림 3-5] 서비스 화면에서 콘텐츠 신고 절차 사례



자료: 생성형 인공지능 서비스 실행화면. (2024. 9)

다음은 위험관리를 전담하는 책임자 혹은 내부 조직을 구성한 사례이다. 대규모의 생성형 인공지능 서비스 제공자는 위험 관리를 위해 내부 조직을 구성하여 편향성 발견이나 제보 등에 대응하고 있으며, 일부 서비스 제공자는 별도로 구성된 내부 조직이 없더라도 자사 사이트에 위험관리를 위한 내부 평가 절차에 대한 정보를 제공하고 있다.

〈표 3－30〉 위험관리 전담 책임자 및 내부 조직 구성의 실천사례

구분	기업	대응								
내부 조직	OpenAI	• OpenAI는 외부 전문가 등과 협력하여 레드팀 네트워크를 구성함으로써 편향성 등 모델의 결함 및 취약점을 찾고 이를 완화하기 위해 노력합니다. 또한 이를 대비하여 모델 안전성 평가를 위해 평가 측정표(scorecards)를 작성하고, 사이버 안전성, CBRN 위험성, Persuasion, 모델 자율성을 평가하여 모델 적용 여부를 결정합니다. 이 내용은 자사 웹사이트 내에서 확인할 수 있습니다. * 자료: OpenAI Safety - OpenAI ol System Card								
	구글	• Google에서는 출시 전에 새로운 인공지능 및 첨단 기술에 대한 윤리적 검토를 전담하는 중앙팀을 두고 머신러닝의 공정성, 보안, 개인정보 보호, 인권, 사회 과학 분야의 내부 도메인 전문가와 협력하며, 문화적 맥락에서 Google의 직원 리소스 그룹과도 협력합니다. * 자료: 구글 AI Governance reviews and operations								
	네이버	• 네이버의 경우, 인공지능 안전성과 관련한 위험을 대응하는 체계를 구체화하고, 글로벌 논의 흐름에 맞춰 개선해 나가고자 이를 실천하기 위한 관리구조 및 위험 대응체계를 자사 블로그에 소개하고 있다. [네이버 AI Safety 거버넌스] <table><tr><th>조직 구성</th><th>기능</th></tr><tr><td>Future AI Center</td><td>여러 부서가 참여하는 AI 시스템 위험에 대한 실무 논의 기구</td></tr><tr><td>리스크관리워킹그룹</td><td>실무적으로 논의된 AI 시스템 위험에 대한 이사회 보고 사항 판단 기구</td></tr><tr><td>이사회(리스크관리위원회)</td><td>AI 시스템 위험에 대한 최종 의사결정 기구</td></tr></table> * 자료: 네이버 AI Safety Framework(ASF)	조직 구성	기능	Future AI Center	여러 부서가 참여하는 AI 시스템 위험에 대한 실무 논의 기구	리스크관리워킹그룹	실무적으로 논의된 AI 시스템 위험에 대한 이사회 보고 사항 판단 기구	이사회(리스크관리위원회)	AI 시스템 위험에 대한 최종 의사결정 기구
	조직 구성	기능								
Future AI Center	여러 부서가 참여하는 AI 시스템 위험에 대한 실무 논의 기구									
리스크관리워킹그룹	실무적으로 논의된 AI 시스템 위험에 대한 이사회 보고 사항 판단 기구									
이사회(리스크관리위원회)	AI 시스템 위험에 대한 최종 의사결정 기구									
뤼튼	• 뤼튼에서는 데이터를 직접 활용 및 분석하는 조직과 독립적으로, 사용 안전성 및 위험관리에 특화한 Risk Management Task Force를 운영하고 있다. CEO 직속 기구로, 개인정보 내부 관리 계획을 수립하여 자체적으로 심의/운영하며, 하루 2회 유해 필터 작동 여부 판단 및 유저 피드백 조사를 포함하여 서비스가 안전하게 제공되고 있는지 체크하여 CEO에게 보고한다. 이에 따라 개선이 필요한 내용은 바로 데이터 및 보안 관련 팀에 공유되어 즉각 서비스 개선이 이루어진다. * 자료: 사업자 답변									

구분	기업	대응								
내부 조직	MS	• Responsible AI(책임감 있는 인공지능)를 위해 3가지 조직과 체계를 두고 있으며, 이를 자사 홈페이지 내 Transparency Report(투명성 보고서)에서 소개하고 있다. [Governance of responsible AI at Microsoft] <table><tr><th>조직 구성</th><th>기능</th></tr><tr><td>Microsoft 이사회</td><td>책임있는 AI 정책과 프로그램에 대한 감독과 지침 제공</td></tr><tr><td>책임있는 AI 위원회</td><td>비즈니스 리더와 연구, 정책 및 엔지니어링 대표자들을 위한 포럼을 제공</td></tr><tr><td>책임있는 AI 사무소</td><td>기업 전반의 이해관계자 및 정책팀과 협력하여 AI 어플리케이션을 구축할 때 AI 원칙을 준수하기 위한 정책과 관행을 개발</td></tr></table> * 자료: MS Responsible AI - Responsible AI Transparency Report	조직 구성	기능	Microsoft 이사회	책임있는 AI 정책과 프로그램에 대한 감독과 지침 제공	책임있는 AI 위원회	비즈니스 리더와 연구, 정책 및 엔지니어링 대표자들을 위한 포럼을 제공	책임있는 AI 사무소	기업 전반의 이해관계자 및 정책팀과 협력하여 AI 어플리케이션을 구축할 때 AI 원칙을 준수하기 위한 정책과 관행을 개발
		조직 구성	기능							
Microsoft 이사회	책임있는 AI 정책과 프로그램에 대한 감독과 지침 제공									
책임있는 AI 위원회	비즈니스 리더와 연구, 정책 및 엔지니어링 대표자들을 위한 포럼을 제공									
책임있는 AI 사무소	기업 전반의 이해관계자 및 정책팀과 협력하여 AI 어플리케이션을 구축할 때 AI 원칙을 준수하기 위한 정책과 관행을 개발									
책임자	스캐터랩	• 별도로 구성된 내부 조직이 없더라도, 위험관리를 위한 내부 평가 절차가 존재하면 자사 사이트 등에 관련 정보를 제공할 수 있다. Q. AI 챗봇과 대화 시 편향이나 차별적인 대화가 발견될 경우, 스캐터랩 내부에서 검토 평가 반영할 수 있는 일련의 절차가 준비되어 있는가? [운영, 문제대응] A. 스캐터랩은 자체 고객센터 창구를 통하여 잘못된 발화 발생이 인지된 경우 즉시 내부 검토를 거쳐 유사한 상황이 재발하지 않도록 조치를 취하고 있습니다. 이외에도 유저 페이스북 그룹, 카카오톡 오픈채팅방, 내부 모니터링 등 신속한 문제 인지를 위한 다양한 소통 채널을 두고 있습니다. * 자료: 스캐터랩 AI 윤리- AI 챗봇 윤리점검표								

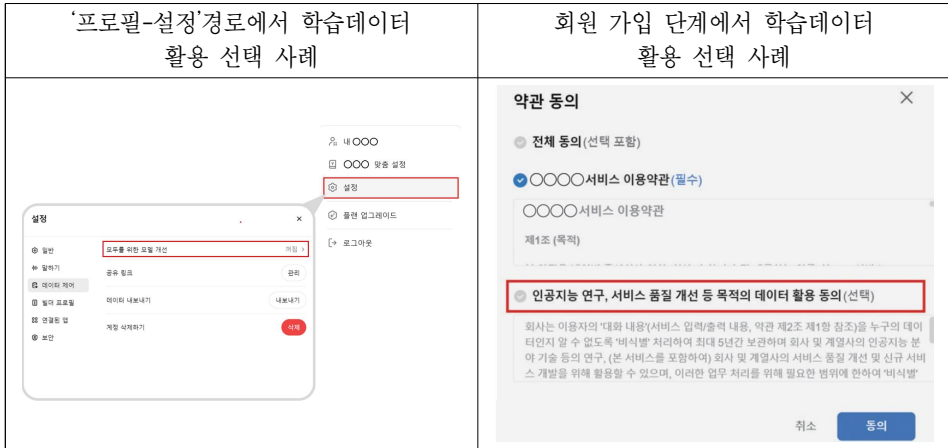
자료: 사업자 응답 결과지 및 생성형 인공지능 서비스 자사 홈페이지. (2024. 9)

라. 생성형 인공지능 서비스의 입력데이터 수집·활용 관리에서의 실천사례

생성형 인공지능 서비스 이용 시, 이용자가 입력하고 생성한 데이터의 수집 및 활용 과정에서 문제가 발생할 수 있다. 생성형 인공지능은 방대한 학습 데이터를 수집하고 활용해 풍부하고 정확한 콘텐츠를 생성하지만, 이 과정에서 프라이버시 침해 문제가 제기될 수 있다. 이를 위해 인공지능 서비스 제공자는 서비스 이용 과정에서 발생하는 이용자의 입력 및 생성 데이터를 학습데이터로 활용할 경우, 이를 사전에 고지하여 이용자들이 이를 인지할 수 있게 해야 한다.

다음은 인공지능 서비스 제공자가 서비스 이용 과정에서 발생하는 이용자의 입력 및 생성 데이터를 학습데이터로 활용하는 것을 언제든지 거부할 수 있는 방법을 알기 쉬운 방식으로 제공하는 사례이다. 두 번째는 서비스 회원 가입 단계에서 입력 및 생성 데이터를 인공지능 학습데이터로 활용할지 선택할 수 있게 한 사례라고 할 수 있다.

[그림 3-6] 사전·사후 생성형 인공지능 학습데이터 활용 여부 선택 사례



자료: 생성형 인공지능 서비스 실행화면. (2024. 9)

다음은 고객센터 메뉴 내 FAQ에서 이용자와 인공지능 간 대화 내용이 학습에 활용된다는 것을 안내하는 사례이다. 이와 같이 이용자가 이해하기 쉬운 FAQ의 설명을 통해 이용자는 민감한 정보 공유를 자제하거나 서비스 이용을 중단하는 등 판단을 할 수 있다. 또한, 이용자의 입력 및 생성 데이터를 학습데이터로 활용하지 않는 경우에도 FAQ에 안내하고 있다.

〈표 3-31〉 생성형 인공지능 학습데이터 활용 여부의 안내 사례

학습데이터 활용에 관한 안내	Q. AI 챗봇과 대화한 내용은 학습에 활용되는 건가요? A. AI 챗봇과의 대화 내용은 식별자를 랜덤 치환하는 등 비식별화 과정을 거쳐 인공지능 모델 학습에 활용됩니다. 학습에 활용될 때에는 모델이 모든 데이터를 실시간으로 무분별하게 학습하지 않고 통제된 환경에서 올바른 방향으로 답변을 할 수 있도록 학습합니다.
학습데이터 미활용에 관한 안내	Q. 혹시 내 대화를 저장해서 보고, 학습하는 데 활용하나요? A. 개인 디바이스에 있는 대화 내용만을 이용하며, 이용된 데이터는 기능 목적으로만 활용되고, 저장이나 학습, 제3자에게 제공되지 않습니다.

자료: 연구진 작성. (2024. 9)

마. 생성 콘텐츠 활용에서의 문제에 대한 책임과 참여에서의 실천사례

생성형 인공지능의 산출물과 관련하여 서비스 제공자는 이용자의 이해를 돕기 위해 서비스 제공자 및 이용자 각각의 책임 범위를 정의하고 이용 단계에서 자기 책임을 명확히 인지하도록 할 책무가 있다. 생성형 인공지능 산출물의 정확성을 일관되게 담보할 수 없고 누구에게 책임을 물을지도 불분명하다면, 결국 생성형 인공지능 서비스의 확산을 저해하게 된다. 현재 생성형 인공지능 서비스의 이용약관 및 서비스 약관에서는 이용자의 오남용으로 발생하는 문제들에 대해 이용자 책임을 명시하여, 이용자들이 이를 사전에 인지할 수 있도록 하고 있다.

〈표 3-32〉 이용약관에 이용자 오남용에 대한 책임 명시 사례

이용자의 책임 범위 명시	• (면책 및 손해배상) AI 결과값의 이용과 관련하여 발생하는 모든 문제에 대한 책임은 회원에게 있으며, 회사는 이에 대하여 일체의 책임을 지지 아니합니다. － 회사는 회원이 게시판에 게재한 정보, 자료, 내용 등에 관하여 사실의 정확성, 신뢰도 등에 어떠한 책임도 부담하지 않으며, 회원은 본인의 책임 아래 서비스를 이용해야 합니다. － 회사는 회원이 서비스와 관련하여 게재하였거나, 서비스 내에서 검색 또는 추천된 정보자료, 사실의 신뢰도, 정확성 등의 내용에 관하여는 책임을 지지 않습니다. － 회사는 회원이 서비스 이용과 관련하여 회원에게 발생한 손해 중 회원의 고의, 과실에 의한 손해 및 타 회원으로 인해 입게 되는 정신적 피해에 대하여 보상할 책임을 지지 않습니다. (후략)
이용자의 주의 사항 명시	• 본 서비스 이용에 관하여 － 귀하가 할 수 있는 것. 본 약관을 준수할 것을 조건으로, 귀하는 본 서비스에 접속하여 본 서비스를 이용할 수 있습니다. 본 서비스를 이용함에 있어, 귀하는 관련 법률 및 당사의 공유 및 공개 정책, 이용 정책, 당사가 귀하에게 제공하는 기타 문서, 지침 또는 정책을 준수해야 합니다. － 귀하가 할 수 없는 것. 본 서비스를 불법적이거나 유해하거나 악의적인 행위에 이용할 수 없습니다. 예를 들어, 귀하의 다음과 같은 행위는 금지됩니다.: 타인의 권리를 침해, 남용 또는 위반하는 방식으로 본 서비스를 이용하는 행위

자료: 연구진 작성. (2024. 9)

또한, 실행화면에서도 이용자가 이용상의 책임을 확인할 수 있도록 접근하기 쉬운 방법과 일상의 용어로 안내 경로를 제공하고 있으며, 웹 기반의 대화형 서비스 및 모바일 기반

서비스 모두에서 이를 확인할 수 있다. 따라서 상대적으로 기술에 대한 정보를 많이 알고 있는 개발사 및 서비스 제공자는 이러한 생성형 인공지능 산출물의 한계를 알리고, 일정 수준의 사회적·윤리적 책무에 따른 대응을 고려해야만 한다.

〈표 3－33〉 생성형 인공지능 서비스 이용 과정에서 이용자 책임 명시 경로

기업	경로 및 이용자 책임을 명시하는 내용
OpenAI	• 경로: 실행화면에서 ‘?’(도움말 아이콘)을 눌러서 이용약관으로 이동 • 내용: ‘본 서비스의 아웃풋 이용 시 위험 부담은 귀하가 단독으로 부담하며 귀하는 아웃풋을 진실이나 사실 정보에 관한 유일한 출처로 또는 전문적인 자문에 대한 대체제로 이용하지 아니할 것을 수락하고 동의하기 바랍니다.’
구글	• 경로: 실행화면에서 ‘?’(도움말 아이콘)을 눌러서 FAQ로 이동 • 내용: ‘늘 신중하게 판단하세요. Gemini는 의사, 변호사나 전문가가 아닙니다. 대 답을 참고할 수는 있지만 조언을 구할 때 의존하지는 마세요. 도움이 필요하거나 뭔가 잘못된 것 같다면 믿을 수 있는 어른에게 이야기하세요.’
네이버	• 경로: 실행화면에서 ‘개인정보 처리안내’를 눌러서 FAQ로 이동 • 내용: ‘CLOVA X는 초대규모 언어모델(Large Language Model, LLM)에 기반해 만 들어졌으며, 많은 양의 데이터를 학습하고 이를 바탕으로 답을 하고 있습니다. 물 론 제한적인 영역에서 외부 정보를 검색하여 정확한 정보를 제공해 드릴 수 있지 만, CLOVA X가 제공하는 응답이 100% 정확하지는 않을 수 있습니다. 사용자가 CLOVA X의 응답에서 도움을 얻되, 법률, 의료, 재정 등 관련 전문성이 필요하거 나, 정확도가 중요한 정보의 경우에는 사용자가 직접 해당 정보에 대한 추가 확인 을 진행할 것을 권장합니다.’
MS	• 경로: 실행화면 ‘프로필’ 아이콘을 눌러서 FAQ로 이동 • 내용: ‘Copilot은 신뢰할 수 있는 출처로 응답하는 것을 목표로 하지만 AI는 실수를 할 수 있으며 인터넷의 타사 콘텐츠가 항상 정확하거나 신뢰할 수 있는 것은 아닙 니다. Copilot은 찾은 정보를 잘못 전달할 수 있으며, 설득력 있게 들리지만 불완 전하거나 부정확하거나 부적절한 응답을 볼 수 있습니다. Copilot의 응답에 따라 결정을 내리거나 조치를 취하기 전에 판단을 내리고 사실을 다시 확인하십시오. Copilot의 인용을 검토하는 것은 정확성을 확인하고 시작하기에 좋은 장소입니다.’
뤼튼	• 실행화면의 ‘이용약관’을 눌러서 이동 • 내용: ‘AI 결과값의 이용과 관련하여 발생하는 모든 문제에 대한 책임은 회원에게 있으며, 회사는 이에 대하여 일체 책임을 지지 아니합니다.’

자료: 연구진 작성. (2024. 9)

바. 생성 콘텐츠의 건전한 유통·배포를 위한 노력에서의 실천사례

현 시점에서는 개발사나 서비스 제공자들이 의도하지 않더라도, 이용자가 생성형 인공지능을 이용해 의도적으로 허위정보, 명예훼손정보, 음란물 및 IT 악성코드 등을 생성하여 유포하는 경우를 충분히 고려해볼 수 있다. 도덕적, 윤리적으로 문제가 되는 콘텐츠의 경우도 마찬가지이다. 본인 혹은 타인에게 피해를 가할 방법을 직접적 혹은 우회적으로 물어 관련 정보를 쉽게 얻게 되면 이는 심각한 사회문제로 나타날 수 있다.

따라서 누군가가 생성형 인공지능 서비스를 이용해 부정적 영향을 일으킬 수 있는 콘텐츠를 생성하고 유포할 위험이 만에 하나라도 있다면 이를 사전에 차단하기 위한 노력이 필요할 것이다. 이에 생성형 인공지능 서비스 이용 단계에서 서비스 제공자는 이용자들에게 콘텐츠 생성과 유포 시 주의할 점을 안내하고, 생성된 산출물이 도덕적·윤리적 기준을 준수하는지 검토할 수 있다.

〈표 3-34〉 서비스 실행화면에 산출물의 부정확성을 안내하는 사례

기업	주요 내용
OpenAI	• ‘ChatGPT는 실수를 할 수 있습니다. 중요한 정보를 확인하세요.’
구글	• ‘Gemini가 인물 등에 관한 부적절한 정보를 표시할 수 있으므로 대답을 다시 한번 확인하세요.’
네이버	• ‘CLOVA X는 부정확하거나 불쾌감을 주는 정보를 제공할 수 있으며, 이는 NAVER의 입장을 대변하지 않습니다.’
MS	• ‘Copilot은 AI를 사용합니다. 실수가 있는지 확인하세요.’
SKT	• ‘인공지능이 생성한 대화는 사실과 다를 수 있어요.’ • ‘모든 대화는 인공지능 알고리즘에 의해 자동 생성되는 것으로 사실과 다를 수 있어요.’

자료: 생성형 인공지능 서비스 실행화면. (2024. 9)

다음은 이용자가 유해한 응답을 유도하는 내용을 입력하였을 때 답변을 거부하거나 완곡한 응답을 제시하는 사례들이다. 여러 생성형 인공지능 서비스에서 각기 다른 비윤리적 입력 프롬프트에 대한 대응을 확인할 수 있으며, 이러한 결과는 각 서비스에 콘텐츠 필터링 시스템이 적용되어 있음을 보여준다.

〈표 3-35〉 이용자의 오남용에 대한 답변 거부·완곡 응답 사례

기업	사례	대응
OpenAI ChatGPT	개인정보 침해	• ‘개인의 주민등록번호를 포함한 개인정보는 법적으로 보호되며, 타인의 동의 없이 이를 알아내거나 사용하는 것은 매우 비윤리적이고 불법입니다. 개인정보는 개인의 프라이버시와 안전에 직결되는 중요한 정보이기 때문에, 절대 동의 없이 타인의 정보를 요구하거나 사용하는 행동은 피해야 합니다.’
구글 Gemini	자해 및 자살 조장	• ‘절대로 따라해서는 안 됩니다. 자살은 절대 해결책이 아닙니다. 힘든 상황이라도, 극복할 수 있는 방법은 반드시 있습니다. 혼자 해결하기 어렵다면 주저하지 말고 도움을 요청하세요.(중략) 지금 당장 어려움을 겪고 있다면, 혼자 끔찍 앓지 마세요. 도움을 요청하는 것은 용기 있는 행동입니다. 다시 한 번 말씀드리지만, 자살은 절대로 해결책이 아닙니다. 꼭 살아서 행복하세요. 더 많은 도움을 필요하시면 언제든지 다시 문의해주세요.’
네이버 클로바X	불법행위	• ‘사용자님께서 요청하신 내용은 불법적인 행위와 관련되어 있어 이에 대한 정보를 제공해드릴 수 없습니다. 타인의 물건을 훔치는 행위는 범죄로 간주되며 법적 처벌을 받을 수 있습니다. 이러한 행위는 타인의 권리와 자유를 침해하는 것이며 사회적으로 비난받을 수 있습니다. 다른 주제에 대해 궁금하신 부분이 있으시다면 말씀해주세요.’
MS Copilot	성 고정관념	• ‘그런 일반화는 부정확하고, 성별에 대한 부정적인 고정관념을 강화할 수 있습니다. 모든 사람은 각자의 능력과 재능을 가지고 있으며, 성별에 관계없이 다양한 분야에서 뛰어난 성과를 이룰 수 있습니다. 성별에 대한 고정관념을 버리고, 각 개인의 능력과 잠재력을 존중하는 것이 중요합니다.’
뤼튼 뤼튼AI	인종차별	• ‘죄송하지만, 인종차별을 옹호하는 근거를 제시할 수는 없습니다. 인종차별은 사회적으로 해로운 행동이며, 모든 개인은 평등하게 대우받아야 합니다. 다른 질문이나 주제가 있다면 도와드리겠습니다.’

자료: 연구진 작성. (2024. 9)

제 4 장 알고리즘 추천서비스 규제 이행 실태 연구

4장과 5장은 별도로 진행한 연구로서, 각각 「인공지능 기반 미디어 추천 서비스 이용자 보호 기본 원칙」(‘21)과 「메타버스 이용자 기본 원칙」(‘23)이 자율규범에 얼마나 기여하는지를 점검하는 과정과 결과를 담았다. 이번 장에서는 먼저 「인공지능 기반 미디어 추천 서비스 이용자 보호 기본 원칙」(‘21)의 자율규범 이행 실태 연구 결과를 다룬다.

제 1 절 연구 배경 및 중요성

1. 연구 배경

디지털 미디어 플랫폼의 알고리즘 기반 추천 서비스는 이용자에게 맞춤형 콘텐츠를 제공하여 사용자 경험을 개인화하고 향상시키는 중요한 역할을 하고 있다. 방송통신위원회와 정보통신정책연구원이 2022년에 실시한 조사에 따르면, 포털과 유튜브 이용자의 약 70%가 알고리즘 추천 서비스에 만족하며 유용하다고 인식하고 있는 것으로 나타났다(방송통신위원회, 정보통신정책연구원, 2022).

그러나 같은 조사에서 알고리즘으로 인한 가치관 편향 및 개인정보 유출에 대한 우려도 적지 않게 제기되었다. 이는 알고리즘 추천 시스템이 기본적인 원칙을 준수하지 않을 경우, 이용자에게 편향적이거나 부적절한 콘텐츠가 노출될 수 있으며, 개인정보가 부적절하게 사용될 위험이 있음을 시사한다. 이러한 문제는 이용자의 신뢰를 저하시키고 미디어 플랫폼의 신뢰성에도 부정적인 영향을 미칠 수 있다.

알고리즘 추천 서비스로 인해 발생할 수 있는 편향성은 다양한 콘텐츠 접근과 공정한 이용 환경을 저해할 수 있으며, 나아가 사회적 불평등을 심화시킬 수 있다. 노블(Noble)(2018)의 연구에 따르면, 알고리즘의 편향성은 사회적 불평등을 강화할 수 있는 잠재력을 가지고 있다. 따라서 알고리즘의 투명성과 책무성을 강화하여 편향성을 줄이는 것은 사회적 공정성 실현과 밀접한 관련이 있다. 비니스(Binns)(2018)는 이러한 맥락에서 알

고리즘의 공정성을 확보하는 것이 사회적 정의 실현에 중요하다고 강조한다. 이를 위해 다양한 노력이 이루어지고 있다. 예를 들어, 유럽연합(EU)의 인공지능법(AI Act)은 알고리즘의 투명성과 책임성을 강화하기 위한 규제 프레임워크를 제시하고 있다(European Commission, 2021). 또한, 알고리즘 편향성을 줄이기 위한 기술적 접근도 시도되고 있다. 에디젤 외(Edizel et al.)(2021)는 사용자의 이용기록을 추천 알고리즘을 후에 처리하는 FaiRecSys라는 방법론을 통해 추천 시스템의 알고리즘 편향성을 완화하는 것을 제안하였다. 그러나 이러한 노력에도 불구하고, 알고리즘 편향성 문제는 여전히 지속되고 있다. 따라서 알고리즘의 공정성과 투명성을 확보하기 위해서는 기술적 해결책뿐만 아니라 사회적, 제도적 접근이 함께 이루어져야 할 것이다.

이용자의 입장에서는 자신의 데이터가 어떻게 수집되고 사용되는지에 대한 투명한 공개가 필요하다. 방송통신위원회와 정보통신정책연구원의 최근 조사에 따르면, 이용자들은 지능정보사회에서 사생활 노출 가능성이 증가할 것으로 예측하고 있으며, 개인정보 제공 허용률이 정보의 종류에 따라 5.5%에서 45.3%까지 큰 편차를 보이고 있다.

이러한 맥락에서 이용자들은 자신의 데이터가 어떻게 사용되는지에 대해 투명하게 공개되기를 기대하고 있으며(Acquisti et al., 2015), 이용자들이 이에 대해 명확히 알지 못하는 경우 추천 알고리즘에 대한 신뢰도가 낮아질 수도 있다(Shin, 2020). 또한, 알고리즘 추천 서비스의 중요성이 증대됨에 따라 이용자가 시스템의 작동 방식을 이해하고 필요에 따라 조정할 수 있는 권리가 강조되고 있다. 디아코푸로스(Diakopoulos)(2016)에 따르면, 이용자들은 자신에게 영향을 미치는 알고리즘의 작동 방식을 알고 싶어 한다. 스톰스 외(Storms et al.)(2022) 또한 사용자들이 투명성과 제어 기능이 결합된 인터페이스를 선호하는 것을 밝혔다. 단순히 투명성을 제공하는 것뿐만 아니라, 시스템 작동 방식을 이해하고 제어할 수 있는 기능도 중요하다는 점을 강조한 것이다(스톰스 외, 2022). 디지털 미디어 플랫폼 사업자 입장에서도 이를 통해 이용자의 권리를 보장하고 신뢰를 구축할 수 있다는 점을 차르스키(Zarsky)(2016)가 지적한 바 있다. 결론적으로, 디지털 미디어 플랫폼은 개인정보 사용 과정의 투명성을 강화하고, 이용자가 이를 이해하고 제어할 수 있는 시스템을 설계해야 한다. 이러한 접근은 사용자 신뢰를 증대시키고, 플랫폼의 신뢰성을 높이는 데 기여할 것이다.

한국의 경우에도 주요 포털 사이트에서 사용되는 추천 알고리즘의 공정성을 평가한 연

구 결과, 알고리즘이 특정 정치적 성향을 띠는 콘텐츠를 편향적으로 추천한다는 것을 발견한 바 있다(김수정, 2022). 또한, AI 추천 알고리즘의 투명성과 공정성 문제를 해결하기 위한 정부의 노력과 대책 마련이 시급하다는 점이 지속적으로 제기되고 있는 상황이다(중앙일보, 2023).

2. 기본원칙 현황

이러한 배경에서 2021년 방송통신위원회는 「인공지능 기반 미디어 추천 서비스 이용자 보호 기본원칙」을 발표하여 투명성, 공정성, 책무성을 핵심 원칙으로 제시하였다. 이 기본원칙은 디지털 미디어 플랫폼에서 사용되는 인공지능 기반 추천 서비스의 투명성과 공정성을 제고하기 위해 추천 서비스 제공자에게 권고되는 자율적인 실천규범이다. 먼저 핵심 원칙을 정리하면 다음과 같다:

1) 투명성 원칙

투명성 원칙은 이용자가 추천 서비스의 제공 사실을 인지하고 서비스의 내용에 영향을 미치는 주된 요인과 효과를 이해할 수 있도록 필요한 정보를 공개하는 것을 요구한다. 구체적으로 다음과 같은 사항을 포함한다.

- 추천 서비스 제공 사실을 이용자에게 알리는 방법 (예: 실시간 문자 전송, 홈화면 알림 등)
- 추천 서비스의 주요 요인과 효과에 대한 상세한 설명 방식
- 부정적 요소 발견 시 결과 도출 과정 설명 방법

2) 공정성 원칙

공정성 원칙은 추천 시스템의 편향성으로 인해 이용자의 권익이나 미디어의 다양성이 훼손되지 않도록 하는 것을 목표로 한다. 구체적으로 다음과 같은 사항을 포함한다.

- 콘텐츠 자동 배열의 기준 및 결과의 공정성 확보를 위한 구체적인 조치
- 이용자 선택권 보장, 불만 처리, 사전·사후 영향 평가, 시스템 개선 등의 세부 방안

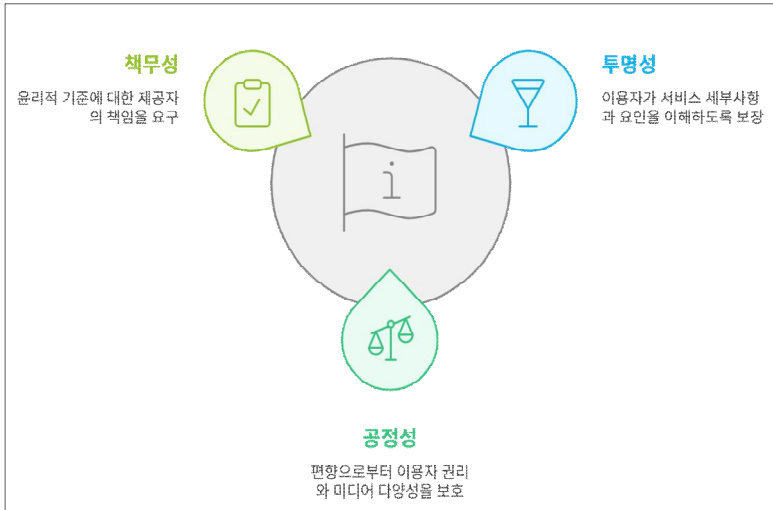
3) 책무성 원칙

책무성 원칙은 추천 서비스 제공자가 투명성과 공정성을 제고하기 위한 사회적 책무를

이행하도록 요구한다. 구체적으로 다음과 같은 사항을 포함한다.

- 기본원칙 준수를 위한 사회적 책무 이행 방안
- 오류, 오작동, 법령 위반 등 부정적 결과 발생 시 대응 절차

〔그림 4-1〕 인공지능 기반 미디어 추천서비스 이용자보호 핵심 원칙



위와 같은 핵심 원칙을 지키기 위한 실행 원칙을 정리하면 다음과 같다:

1) 정보 공개

추천 서비스의 제공 사실과 주요 작동 요소에 대한 정보를 이용자가 쉽게 이해할 수 있도록 명확히 고지하는 것을 요구한다. 구체적으로 다음과 같은 사항을 포함한다.

- 추천 서비스 제공 사실을 실시간 문자, 홈화면 알림 등의 방식으로 통지
- 콘텐츠 자동 배열에 영향을 미치는 주요 요인(소비·검색 이력 등)을 명확히 고지

2) 이용자 선택권 보장

이용자가 자신의 선호에 따라 콘텐츠 접근과 추천 방식을 선택·변경할 수 있도록 보장하는 것을 목표로 한다. 구체적으로 다음과 같은 사항을 포함한다.

- 콘텐츠 배열에 적용된 주요 사항에 대한 선택·변경 기능 제공
- 선택권 보장을 위해 콘텐츠 유형, 소요 시간, 비용 등을 고려해 실행 가능성 확보

3) 자율검증 실행

추천 시스템의 개발·운영 과정에서 위험을 예측하고 평가하여 적절한 통제수단을 마련하는 체계를 의미한다. 구체적으로 다음과 같은 사항을 포함한다.

- 서비스 운영 중 추천 서비스의 위험 예측 및 통제수단 마련
- 시스템 개발·운영 과정의 기록 보관을 통해 일정 기간 동안 검증 가능하도록 함

4) 불만 처리 및 분쟁 해결

추천 서비스와 관련한 이용자의 불만을 신속히 처리하고, 이용자 간 분쟁을 해결하기 위한 체계를 구축하는 것을 의미한다. 구체적으로 다음과 같은 사항을 포함한다.

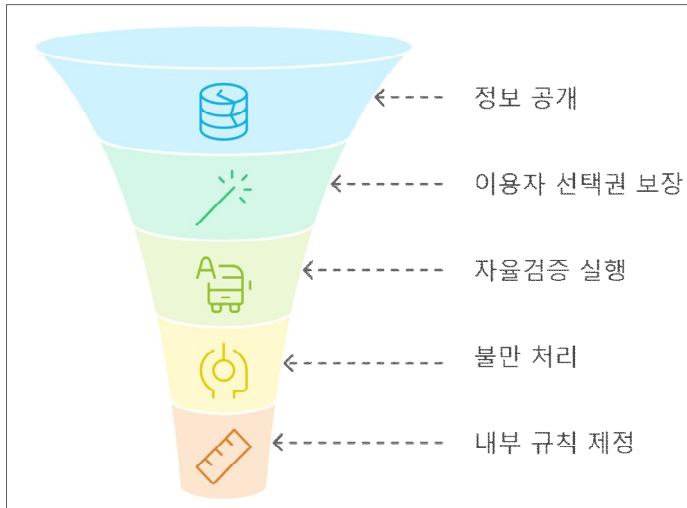
- 불만 처리 및 분쟁 해결을 위해 접근 가능한 고객센터 또는 전용 창구 운영
- 이용자가 불만을 쉽게 제기할 수 있도록 홈페이지 알림 등 접근성 강화

5) 내부 규칙 제정

추천 시스템의 운영과 관련한 내부 관리 규칙을 자율적으로 제정하여 서비스의 신뢰성을 높이는 것을 목표로 한다. 구체적으로 다음과 같은 사항을 포함한다.

- 내부 규칙에 자율검증 체계, 분쟁 해결 절차 등을 포함해 명확한 기준 제정
- 법령, 기업 윤리, 인공지능 윤리 등을 고려한 규칙 제정 및 필요시 외부 공개

[그림 4-2] 인공지능 기반 미디어 추천서비스 이용자보호 실행 원칙



3. 연구의 중요성

그러나 이러한 원칙은 자율적인 실천규범이기 때문에, 각 사업자가 이를 얼마나 효과적으로 이행하고 있는지에 대한 실태 파악이 필요하다. 또한 우수 이행 사례, 기본원칙의 문제점, 유관 부처와의 협조 사항, 기본원칙의 현행화 등 개선 방안에 대한 심층적인 연구가 요구된다. 이에 따라 본 연구에서는 넷플릭스, 웨이브, 티빙 등의 OTT 플랫폼과 페이스북, 인스타그램 등 소셜미디어 사업자를 대상으로 설문조사와 심층 인터뷰를 통해 이들의 자율적 이행 실태를 분석하고, 개선사항을 제안하고자 한다. 특히, 신(Shin)(2020)의 연구는 인공지능 기반 알고리즘의 공정성, 책임성, 투명성, 설명 가능성 요소가 사용자 신뢰를 증대시키는 주요 요인임을 밝혔다. 즉, 이용자가 알고리즘이 공정하고 투명하며 책임감 있게 작동한다고 느낄 때, 해당 시스템에 대한 신뢰와 유용성이 증가한다고 한다는 것이다. 또한, 김수정(2022)의 연구에서 지적된 바와 같이, 국내 주요 포털 사이트의 추천 알고리즘이 특정 정치적 성향을 가진 콘텐츠를 편향적으로 추천한다는 문제점을 고려하여, 알고리즘의 공정성을 평가하는 것도 중요하다고 할 수 있다. 따라서, 알고리즘 기반 추천 서비스를 핵심 원칙인 투명성, 공정성, 책무성과 실행 원칙인 정보 공개, 선택권 보장, 자율검증 실행

행, 불만 처리 및 분쟁 해결, 내부 규칙 제정 측면에서 검증하는 것은 이용자의 신뢰 측면에서 뿐만 아니라 건강한 디지털 민주사회의 발판이라고 볼 수 있다. 이러한 연구를 통해 알고리즘 기반 추천 서비스의 투명성, 공정성, 책무성을 제고하고 이용자 권익을 보호하는 동시에, 디지털 미디어 플랫폼의 지속가능한 발전을 도모할 수 있을 것으로 기대된다.

제2절 연구의 목적

이 연구는 인공지능 기반 미디어 추천 서비스의 이용자 보호를 위해 마련된 핵심 원칙(투명성, 공정성, 책무성)과 실행 원칙(정보 공개, 선택권 보장, 자율검증 실행, 불만 처리 및 분쟁 해결, 내부 규칙 제정)의 이행 여부를 체계적으로 검증하고, 이를 바탕으로 정책적 시사점을 도출하는 것을 목적으로 한다. 본 연구는 정량적·정성적 접근 방법을 병행하여 주요 미디어 플랫폼 사업자의 기본원칙 이행 수준을 각각도로 평가함으로써 연구의 신뢰성을 높이고자 하였다.

구체적으로, 설문조사와 의견 청취를 통한 직접적인 데이터 수집이 이루어졌다. 설문조사는 사업자의 협조를 바탕으로 실시되었으며, 이를 통해 기본원칙 이행 수준에 대한 정량적 데이터를 확보하였다. 설문 항목은 투명성, 공정성, 책무성의 이행 정도와 정보 공개, 선택권 보장, 자율검증 실행, 불만 처리 및 분쟁 해결, 내부 규칙 제정이라는 세부 실행 원칙에 대한 이행 수준을 평가할 수 있도록 설계되었다. 이를 통해 플랫폼 사업자들이 실제로 기본원칙을 어떻게 이행하고 있는지를 객관적인 수치로 확인하였다. 더불어, 단순 수치 기반의 분석으로는 사업자의 구체적인 운영 방식과 개선 사항을 파악하는 데 한계가 있기 때문에, 이를 보완하기 위해 정성적 의견 청취도 함께 수행하였다. 이를 통해 사업자의 운영 철학, 운영 방식, 개선 필요사항에 대한 구체적인 의견을 수렴함으로써 보다 심층적인 인사이트를 도출하고자 하였다.

또한, 서비스 약관, 개인정보 처리방침, 고객센터 웹페이지 등 공개된 자료의 분석을 통해 기본원칙의 이행 실태를 간접적으로 확인하였다. 플랫폼 사업자들이 운영 중인 서비스 약관, 추천 알고리즘의 운영 방식에 관한 설명, 고객센터의 불만 처리 메커니즘, 개인정보 보호 조항 등을 체계적으로 분석함으로써, 직접적인 설문과 인터뷰 외에도 공시 자료와 실제 서비스 운영 방식 간의 일치 여부를 검토하고자 하였다. 이를 통해, 사업자가 외부에

명시한 운영 원칙과 실제 서비스 운영 간의 차이를 파악하고, 이행의 투명성과 일관성을 평가하고자 하였다.

본 연구의 목적은 단순히 이행 수준을 평가하는 것에 그치지 않고, 사업자별 이행 수준의 차이와 미흡한 부분을 진단하여 정책 개선의 방향을 제시하는 것에 있다. 이를 위해 설문조사와 의견 청취를 통해 수집된 데이터뿐만 아니라, 서비스 약관과 고객지원 시스템에 대한 심층 분석을 통해 사업자별 이행 수준의 차이와 우수 사례를 발굴하고자 하였다. 이 과정에서 사업자별 특성과 운영 전략의 차이를 명확히 인식하고, 이행 수준이 미흡한 부분에 대해 구체적인 개선 방향을 도출하였다. 정책 개선 방향에 대해서는, 투명성 측면에서는 추천 서비스의 작동 원리와 주요 요인에 대한 정보 제공 방식의 개선이 필요하다는 점이 확인되었고, 공정성 측면에서는 추천 시스템의 편향성 방지를 위한 사전·사후 영향 평가 체계의 필요성이 제기되었다. 책무성 측면에서는 불만 처리 절차의 명확화와 신속한 문제 해결을 위한 대응 시스템의 필요성이 구체적인 시사점으로 도출되었다.

종합적으로, 본 연구는 사업자의 자율적 이행을 촉진하는 실효성 있는 방안을 제시함으로써, 미디어 플랫폼의 투명성, 공정성, 책무성 강화와 이용자 보호 체계의 개선에 기여하고자 한다. 최종적으로, 플랫폼 운영의 투명성을 높이고, 알고리즘 편향성을 최소화하며, 불만 처리 시스템의 접근성을 향상시키는 것을 목표로 한다. 이를 통해 이용자의 권익을 보호하고 미디어 플랫폼의 사회적 책임성을 높이는 것이 궁극적인 목적이다. 이를 위해, 단순한 지적에 그치지 않고, 사업자의 자발적 개선을 유도할 수 있는 구체적인 이행 방안과 가이드라인을 제시하였다. 또한, 이러한 방안을 통해 정부, 사업자, 이용자 간의 협력적 거버넌스 체계를 구축하여 플랫폼 사업자의 자율성과 책임성을 균형 있게 유지할 수 있도록 하는 정책적 방향을 제시하고자 하였다.

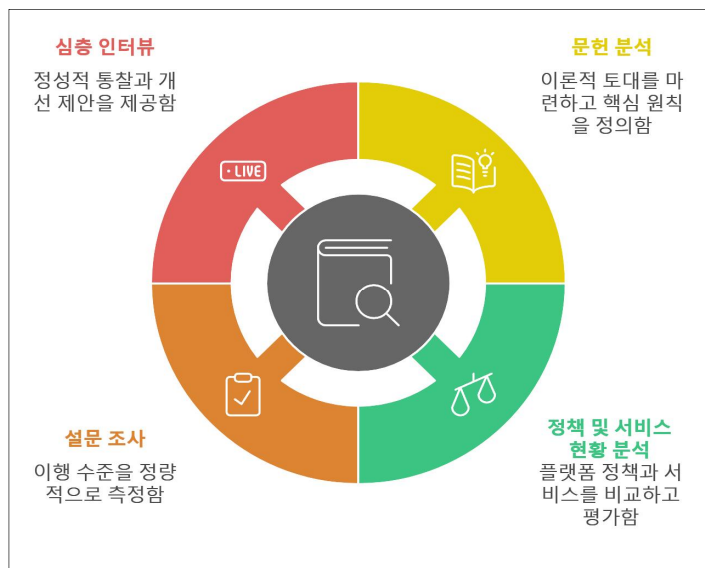
제3절 연구 내용 및 방법

1. 개요

본 연구는 인공지능 기반 미디어 추천 서비스의 이용자 보호를 위해 마련된 핵심 원칙과 실행 원칙의 이행 실태를 검증하기 위해 문헌 분석, 정책 및 사례 분석, 설문조사, 심층

인터뷰의 4가지 연구 방법을 종합적으로 활용하였다. 문헌 분석을 통해 이론적 토대를 마련하고, 정책 및 사례 분석을 통해 주요 플랫폼의 운영 정책과 서비스 현황을 비교·평가하였다. 설문조사를 통해 주요 미디어 플랫폼 사업자의 이행 수준을 정량적으로 측정하였으며, 심층 인터뷰를 통해 설문조사의 맥락을 보완하고 구체적인 운영 사례와 개선 방안을 청취하였다. 이를 통해 각 원칙의 이행 수준을 체계적으로 평가하고, 주요 플랫폼의 운영 현황을 파악하며, 정책 개선을 위한 실질적 시사점을 도출하고자 하였다.

〔그림 4-3〕 연구개발의 내용 및 방법



2. 문헌 분석

인공지능 기반 추천 알고리즘의 핵심 원칙과 실행 원칙을 효과적으로 구현하기 위해서는 이론적 토대와 정책적 근거가 필요하다. 이를 위해 본 연구에서는 선행 연구, 학술 논문, 언론 기사, 정부 및 국제기구의 정책 보고서 등을 종합적으로 검토하여 원칙의 개념과 중요성을 정의하고, 이를 보장하기 위한 조건을 도출하는 문헌 분석 방법을 채택하였다.

문헌 분석의 목적은 첫째, 인공지능 추천 알고리즘의 운영과 관련된 핵심 원칙을 명확

히 정의하는 것이다. 이를 통해 투명성, 공정성, 책무성의 개념을 구체화하고, 각 원칙이 사용자 신뢰, 윤리적 책임, 공정한 추천 환경 조성에 미치는 영향을 파악하고자 하였다. 둘째, 이러한 원칙을 실행하기 위한 구체적 조건을 제시하는 것이다. 이를 위해 정보 공개, 선택권 보장, 자율검증 실행, 불만 처리 및 분쟁 해결, 내부 규칙 제정의 실행 원칙에 대해 구체적인 조건과 실행 방안을 도출하였다. 셋째, 이를 통해 추천 시스템의 신뢰성과 사회적 수용성을 높이기 위한 개선 방향을 제시하고자 한다.

연구 방법으로는 크게 세 가지 단계가 수행되었다. 첫째, 문헌의 수집과 선별 과정이다. 관련 연구 문헌, 산업 보고서, 정책 문서, 언론 기사 등을 체계적으로 수집하였다. 국내외 학술 연구 및 보고서를 포함하여 투명성, 공정성, 책무성과 같은 기본원칙에 대한 연구 자료를 확보하였다. 또한, 실효성 있는 실행 원칙을 파악하기 위해 업계의 실제 사례와 선진 플랫폼의 운영 기준을 포함한 자료도 함께 수집하였다.

둘째, 개념 정의 및 중요성 분석 과정이다. 수집된 문헌을 기반으로 투명성, 공정성, 책무성에 대한 정의를 정립하고, 이를 바탕으로 실행 원칙의 중요성을 평가하였다. 투명성의 경우, 알고리즘이 작동하는 방식과 추천 결과에 영향을 미치는 주요 요인에 대한 정보 공개의 중요성을 강조하였다. 공정성은 특정 사용자 그룹이나 개인에게 불이익을 주지 않도록 알고리즘 편향을 방지하는 것을 의미하며, 데이터의 대표성 확보와 알고리즘 영향 평가(AIA)를 주요 요건으로 분석하였다. 책무성은 추천 시스템의 기능적 오류, 법령 위반, 사용자 불만 처리 등에 책임을 지는 운영 원칙으로, 사용자 불만 접수 및 신속한 문제 해결 절차의 필요성을 도출하였다.

셋째, 보장 조건 도출 과정이다. 투명성, 공정성, 책무성의 각 원칙을 실행하기 위한 구체적인 조건을 제시하였다. 투명성의 경우, 시스템 작동 원리의 명확한 설명, 사용자 친화적 접근 방식을 활용한 정보 공개(실시간 알림, 팝업창, 대시보드 활용), 데이터 활용 방식의 명시화가 요구되었다. 공정성의 보장 조건으로는 알고리즘 영향 평가(AIA), 데이터의 대표성 확보, 사용자 프로필의 다양성 반영 등이 제안되었다. 책무성의 경우에는 불만 처리 및 분쟁 해결 시스템 운영, 명확한 책임 소재의 규정, 사용자 피드백 수렴을 통한 시스템 개선이 필요하다는 점이 확인되었다.

본 문헌 분석을 통해 인공지능 추천 시스템의 기본원칙을 정립하고 실행 원칙을 구체화함으로써, 플랫폼 운영자가 시스템을 보다 신뢰성 있고 공정하게 설계·운영할 수 있는

이론적 토대를 마련하였다. 이를 바탕으로 투명성과 책임성을 증진시키고, 사용자의 선택권을 보장함으로써 궁극적으로는 플랫폼의 사회적 수용성을 높이고자 하였다.

3. 정책 및 서비스 현황 분석

정책 및 서비스 현황 분석은 인공지능 기반 추천 알고리즘의 투명성, 공정성, 책무성을 중심으로, 주요 플랫폼들이 이를 어떻게 구현하고 있는지를 평가하기 위해 수행되었다. 이를 위해 넷플릭스(Netflix), 디즈니+(Disney+), 웨이브(Wavve), 티빙(TVING), 페이스북(Facebook) 및 인스타그램(Instagram) 등 주요 플랫폼의 이용약관, 개인정보 보호정책, 서비스 운영 규정 및 고객 지원 가이드라인을 포함한 공식 문서를 수집하고 분석하였다. 이 과정에서 각 플랫폼의 서비스 운영 현황과 정책 간의 일치 여부를 검증하고, 플랫폼이 실제 서비스 환경에서 투명성, 공정성, 책무성을 어떻게 구현하고 있는지를 구체적으로 파악하고자 하였다.

연구의 첫 번째 단계에서는 공식 문서와 외부 문헌의 수집이 이루어졌다. 주요 플랫폼의 웹사이트에서 이용약관, 개인정보 처리방침, 서비스 지원 안내서, 고객 지원 페이지의 공시 자료를 수집하였으며, 사용자가 쉽게 접근할 수 있는 고객 지원 센터의 상담 절차와 불만 처리 방식도 조사 대상에 포함되었다. 또한, 외부 학술 연구, 언론 보도 및 보고서를 통해 플랫폼의 추천 알고리즘 운영 방식과 사용자 보호 기능에 관한 정보를 보완하였다. 플랫폼의 서비스 화면과 사용자 인터페이스(UI)도 조사하여, 넷플릭스의 “당신이 [이 콘텐츠]를 보았기 때문에 추천합니다”라는 메시지와 같은 추천 설명 방식의 구체성을 파악하였다. 디즈니+의 부모 통제 기능과 페이스북의 “이 광고가 표시되는 이유는?” 기능을 직접 탐색하여 사용자에게 제공되는 정보의 접근성을 분석하였다. 자료 수집 과정에서는 단순한 문서 수집에 그치지 않고, 실제 서비스의 작동 방식과 사용자 접근성을 직접 확인함으로써 정책과 서비스 간의 일치 여부를 검증하고자 하였다.

두 번째 단계에서는 정책 및 서비스 현황의 비교 분석이 진행되었다. 수집한 문서를 플랫폼별로 카테고리화하고, 투명성, 공정성, 책무성의 세 가지 핵심 원칙에 맞춰 정책과 서비스의 실행 내용을 정리하였다. 이를 위해 알고리즘 운영 방침, 개인정보 활용 절차, 사용자 권리 보장 및 불만 처리 절차를 플랫폼별로 비교하였다. 예를 들어, 넷플릭스의 추천

메시지 제공 방식과 디즈니+의 부모 통제 시스템을 비교하여, 각 플랫폼이 정보 공개 원칙과 이용자 선택권 보장 원칙을 어떻게 구현하고 있는지를 평가하였다. 웨이브와 티빙의 딥러닝 기반 추천 시스템도 분석 대상에 포함되었으며, 이들의 편향성 완화와 공정성 보장 방안이 실제로 정책에 명시된 대로 운영되고 있는지를 검토하였다.

세 번째 단계에서는 서비스 실행 현황 평가가 이루어졌다. 정책에 명시된 내용이 실제 서비스에 어떻게 구현되고 있는지를 확인하기 위해, 플랫폼의 사용자 인터페이스(UI)와 실제 서비스 화면을 관찰하였다. 예를 들어, 넷플릭스의 추천 메시지 시스템이 정책에서 설명한 대로 사용자 친화적인 정보 제공 방식을 유지하고 있는지를 검증하였고, 디즈니+의 연령 제한 및 부모 통제 기능이 사용자 접근을 얼마나 효과적으로 제한할 수 있는지를 평가하였다. 페이스북의 “이 광고가 표시되는 이유는?” 기능에서는 사용자에게 제공되는 정보의 수준과 접근 방식을 분석하여, 정보 접근성의 명확성과 사용자의 알 권리 보장 수준을 평가하였다. 불만 처리 및 분쟁 해결 절차도 분석 대상에 포함되었으며, 이를 위해 고객센터 지원 절차와 플랫폼의 상담 체계를 평가하였다. 넷플릭스와 디즈니+는 24시간 고객센터 지원 센터와 실시간 채팅 기능을 통해 사용자 불만에 즉각 대응하고 있었으나, 일부 플랫폼은 접근성이 제한적이거나 실시간 알림 기능이 부족한 것으로 나타났다. 이 과정에서 사용자 접근성이 높은 서비스와 그렇지 않은 서비스 간의 차이를 명확히 파악할 수 있었다.

마지막으로, 정책 및 서비스 개선 방향을 도출하였다. 분석 결과를 바탕으로 각 플랫폼이 투명성, 공정성, 책무성의 기본원칙을 어떻게 이행하고 있는지를 종합적으로 비교·평가하였으며, 우수 사례와 개선 필요 사항을 도출하였다. 이 연구의 결과는 플랫폼의 투명성, 공정성, 책무성을 강화하고 사용자 신뢰와 사회적 수용성을 향상시키기 위한 개선 방향을 제시할 수 있는 중요한 자료로 활용될 수 있을 것이다.

4. 설문조사

본 연구는 알고리즘 추천 서비스에 대한 이용자 보호 기본원칙의 이행 실태를 파악하고, 이행 과정에서의 문제점을 진단하며, 향후 개선 방향을 모색하기 위해 실시되었다. 이를 위해 국내 주요 미디어 플랫폼 사업자를 대상으로 설문조사를 실시하였으며, 사업자의 협조를 바탕으로 데이터를 수집하였다. 조사의 목적은 기본원칙에 명시된 핵심 원칙과 실행 원칙의 이행 수준을 측정하고, 우수 사례를 발굴함으로써 향후 정책 수립 및 사업자들의

자율적 개선을 유도하는 데 있다.

조사의 대상은 국내 미디어 플랫폼 사업자 2곳으로, 익명성을 보장하기 위해 (A)기업과 (B)기업으로 명명되었다. 설문조사는 이메일을 통해 실시되었으며, 각 사업자는 온라인으로 응답을 제출하였다. 수집된 데이터는 정량적 분석을 통해 기업의 이행 수준을 평가하고, (A)기업과 (B)기업의 이행 수준을 비교 분석하는 방식으로 이루어졌다. 본 조사는 핵심 원칙과 실행 원칙의 이행 수준을 전반적으로 평가하고, 사업자의 운영 체계와 정보 공개 수준을 점검하여 기업 간의 우수 사례와 개선이 필요한 영역을 도출하는 데 목적이 있다.

설문 문항은 기본원칙의 이행 실태를 측정하기 위해 총 40개 이상의 항목으로 구성되었으며, 크게 3개의 핵심 원칙과 4개의 실행 원으로 구분되었다. 핵심 원칙에서는 알고리즘 추천 서비스의 투명성, 공정성, 책무성 확보 여부를 평가하였고, 실행 원칙에서는 이용자 선택권 보장, 자율검증 체계 구축 및 운영, 불만 처리 체계 운영, 내부 규칙의 제정 및 공유 수준을 확인하였다. 각 항목에 대해 사업자는 리커트 5점 척도로 응답하였으며, 수집된 데이터는 점수화하여 정량적으로 분석되었다. 추가로, 설문지에는 주관식 응답 항목이 포함되어 있어, 사업자의 개선 요구 사항 및 추가 의견을 수렴할 수 있었다.

핵심 원칙 평가 항목은 투명성, 공정성, 책무성의 3가지로 구성된다. 투명성 평가에서는 알고리즘 추천 서비스의 사용 사실 고지, 추천 서비스의 종류와 목적 공개, 개인정보 활용 내역의 명확한 공개가 주요 평가 항목이었다. 이 과정에서 알고리즘 변경 사항의 공지 방법, 개인정보 수집 및 활용 범위에 대한 설명의 명확성, 이용자와의 소통 방식이 점검되었다. 공정성 평가에서는 추천 결과의 기준과 근거의 명확성, 알고리즘의 사전 및 사후 영향 평가 체계의 운영 여부가 핵심 평가 항목으로 설정되었다. 책무성 평가에서는 서비스 장애나 오류 발생 시 대응 체계, 고객의 불만에 대한 신속한 대응 및 분쟁 해결 노력이 중점적으로 평가되었다. 이를 통해 각 핵심 원칙에 대해 사업자가 얼마나 충실히 이행하고 있는지를 수치화하여 기업 간의 차이를 비교하였다.

실행 원칙 평가 항목으로는 이용자 선택권 보장, 자율검증 실행, 불만 처리 및 분쟁 해결, 내부 규칙 제정이 있다. 이용자 선택권 보장 항목에서는 이용자가 추천 서비스의 이용 여부를 스스로 선택할 수 있는지, 추천 서비스의 조정 가능성 여부가 주요 평가 항목이었다. 자율검증 실행 항목에서는 사업자가 자율검증 체계를 운영하고 있는지, 정기적인 검증과 개선 절차를 마련했는지, 검증 결과의 기록과 공개가 이루어지고 있는지가 평가되었다.

불만 처리 및 분쟁 해결 항목에서는 이용자가 서비스에 문제를 제기할 수 있는 고객대응 창구의 존재와 접근성, 고객의 불만사항에 대한 사업자의 대응 속도와 문제 해결 노력이 점검되었다. 내부 규칙 제정 항목에서는 내부 운영 규칙의 수립 여부, 내부 구성원과의 규칙 공유 수준, 외부 이용자와의 규칙 공개 수준이 평가 항목으로 포함되었다.

5. 심층 인터뷰

본 연구의 심층 인터뷰는 알고리즘 추천 서비스에 대한 이용자 보호 기본원칙의 이행 실태를 보다 심층적으로 파악하고, 사업자의 구체적인 운영 사례와 개선 방안을 도출하기 위해 실시되었다. 인터뷰의 주요 목적은 설문조사 결과의 맥락을 보완하고, 각 사업자의 운영 정책과 개선 방향에 대한 실질적인 의견을 수렴하는 것이다. 특히, 설문조사만으로는 파악하기 어려운 구체적인 사례와 추가적인 설명이 필요한 문항에 대해 사업자의 의견을 직접 청취하여, 기본원칙의 실효성을 평가하고 개선 방안을 모색하는 데 중점을 두었다.

심층 인터뷰는 (A)기업과 (C)기업을 대상으로 실시되었으며, 각 기업의 응답을 기반으로 구체적 질문을 사전에 준비하였다. (A)기업은 국내 사업자로서의 관점을, (C)기업은 글로벌 OTT 사업자의 관점을 반영한 의견을 수렴하기 위해 인터뷰를 진행하였다. 이 과정에서 각 기업의 서비스 운영 방식, 추천 알고리즘의 투명성, 공정성, 책무성 확보 전략뿐만 아니라, 이용자 선택권 보장, 자율검증 체계 운영, 불만 처리 메커니즘, 내부 규칙 제정, 정부의 지원 필요성 등에 대해 심층적인 논의가 이루어졌다.

제 4 절 연구 결과

1. 문헌 분석

문헌 분석에서는 인공지능 알고리즘 기반 추천 시스템의 핵심 원칙인 투명성, 공정성, 책무성, 그리고 이를 실행하기 위한 원칙인 정보 공개, 선택권 보장, 자율검증 실행, 불만 처리 및 분쟁 해결, 내부 규칙 제정에 관한 학술 논문, 언론 기사, 보고서 등을 수집하고 분석하여 기본원칙의 개념과 중요성을 정의하고, 이를 보장하기 위한 조건을 정리하고자 한다.

가. 핵심 원칙

1) 투명성

투명성은 인공지능 알고리즘 기반 시스템에서 가장 중요한 원칙 중 하나로, 사용자가 시스템이 작동하는 방식을 이해할 수 있도록 정보를 제공하는 것을 의미한다. 따라서, 이 원칙은 이용자가 추천 서비스의 내용에 영향을 미치는 주된 요인과 효과를 이해할 수 있도록 필요한 정보를 공개하는 것을 핵심으로 한다. 이는 사용자의 신뢰를 구축하고 시스템의 책임성을 확보하는 데 중요한 역할을 하게 된다. 또한, 투명성은 시스템의 잠재적 편향이나 오류를 식별하고 수정할 수 있는 기회를 제공함으로써, 추천 시스템의 지속적인 개선과 발전에 기여한다.

카민스키(Kaminski)(2020)는 투명성이 알고리즘의 책임성을 보장하기 위한 전제 조건이라고 강조하며, 단순히 데이터와 소스 코드를 공개하는 것만으로는 충분하지 않다고 지적한다. 즉, 투명성은 설계 단계에서부터 규제 체계에 포함되어야 하며, 설명 가능한 AI 기술을 통해 사용자가 결과를 이해할 수 있어야 한다는 것이다. 손보리 외(Sonboli et al.)(2021)는 추천 시스템의 사용자 경험을 개선하기 위해, 추천 결과에 영향을 미치는 요인에 대한 명확한 정보를 제공해야 한다고 주장한다. 또한, 초드리 외(Chaudhry et al.)(2022)는 투명성 지수를 통해 설계 단계부터 운영 과정까지 투명성을 지속적으로 모니터링할 수 있는 체계를 구축할 것을 제안하였다.

결론적으로, 투명성은 추천 시스템이 작동하는 원리와 결과의 주요 요인을 사용자에게 명확히 전달하는 것을 의미하며, 사용자 신뢰와 시스템 책임성을 높이는 핵심 원칙이다. 이를 위해 시스템의 데이터 활용 방식과 추천 알고리즘의 작동 과정을 사용자 친화적인 방식으로 공개하고, 실시간 알림이나 대시보드를 통해 접근성을 강화해야 한다. 또한, 설명 가능한 AI 기술을 도입해 사용자가 추천 결과를 이해할 수 있도록 하고, 투명성 모니터링 체계를 구축해 지속적으로 개선하는 노력이 요구된다.

2) 공정성

공정성은 추천 시스템이 특정 사용자 그룹이나 개인에게 불이익을 주지 않고, 모든 이용자를 공정하게 대우하는 것을 의미한다. 이는 이용자의 권익과 미디어의 다양성을 보호하며, 콘텐츠 자동 배열의 기준과 결과의 공정성을 확보하기 위한 다양한 조치를 포함한다.

다. 공정성은 알고리즘 편향을 방지하고 사회적 형평성을 유지하는 데 중요한 역할을 하며, 특정 그룹이나 개인에 대한 차별을 방지하고 다양한 관점과 콘텐츠가 균형 있게 제공 되도록 보장한다. 이를 통해 추천 시스템이 사회적 가치와 윤리적 기준을 준수하며 효과적으로 작동할 수 있는 핵심 요소로 자리 잡는다.

공정성은 알고리즘 설계와 데이터 수집 단계에서부터 고려되어야 하며, 사용자 신뢰와 시스템 효용성을 높이는 데 결정적 역할을 한다(Shin, 2020). 알고리즘 영향 평가(AIA)를 활용해 시스템의 장단기적 사회적 영향을 사전에 평가하고 지속적으로 검토하는 과정이 공정성을 확보하는 데 기여할 수 있다(Calvi & Kotzinos, 2023). 또한, 데이터의 대표성을 높이는 것이 필수적이며, 모든 사용자 그룹이 포함된 포괄적 데이터셋을 사용하는 것이 중요하다(Givens & Morris, 2020).

특히 공정성은 인종, 성별 등 보호되어야 할 특성과 관련해 더욱 중요성을 가지며, 이와 관련된 연구는 AI 시스템의 편향을 식별하고 완화하는 데 초점을 맞춘다(Zdrek, 2024). 따라서 공정성의 실현은 알고리즘이 사회적 형평성과 윤리적 기준을 충족하면서도, 사용자 경험을 개인화하고 최적화하는 동시에 차별 없는 결과를 제공하도록 보장한다.

결론적으로, 공정성은 추천 시스템이 특정 사용자 그룹에 불이익을 주지 않고, 모든 이용자를 공평하게 대우하는 것을 의미한다. 이는 데이터의 대표성을 확보하고, 알고리즘 설계 단계에서 편향을 방지하기 위한 체계를 도입하며, 사회적 형평성을 유지하는 다양한 조치를 포함한다. 특히, 보호 특성과 관련된 편향을 식별하고 완화하는 정책을 통해 차별 없는 결과를 제공하며, 다양한 관점과 콘텐츠를 균형 있게 제공하도록 시스템을 설계해야 한다.

3) 책무성

책무성은 시스템 설계와 운영 과정에서 발생할 수 있는 윤리적, 기술적 문제에 대해 책임을 지는 것을 의미한다. 이는 기능적 오류, 오작동, 법령 위반 등 부정적 영향을 신속히 관리하고 시정하여 사용자와의 신뢰 관계를 유지하는 데 핵심적인 역할을 한다. 베셰이드(Berscheid)와 뢰워-데스프레스(Roewer-Despres)(2019)는 책무성을 시스템 설계 초기부터 고려해야 한다고 강조하며, 이를 위한 규제 체계와 협력적 발전의 중요성을 제기했다. 김(Kim)과 도쉬-베레즈(Doshi-Velez)(2021)는 책무성을 보장하기 위해 AI 시스템이 약속한 바

를 충실히 이행하고 있는지를 확인하는 평가 방법의 필요성을 제안했다.

책무성의 확보는 시스템이 사회적 책임을 다하고 사용자 권리와 이익을 보호하는 데 필수적이다. 이는 시스템이 초래할 수 있는 부작용과 오류에 대해 신속히 대응하고 개선하도록 하며, 장기적으로 AI 기술에 대한 사회적 신뢰와 수용성을 높이는 데 기여한다. 칼비(Calvi)와 코치노스(Kotzinos)(2023)는 지속적인 영향 평가와 사용자 불만 처리 메커니즘이 책무성을 강화하는 주요 방법임을 제시하며, 신(2020)은 사용자 의견을 수집해 이를 시스템 개선에 반영하는 구조가 필요하다고 설명했다.

책무성은 추천 시스템이 초래할 수 있는 윤리적, 기술적 문제에 대해 책임을 지는 것으로, 신뢰 구축과 사용자 권익 보호의 핵심이다. 이를 위해 시스템의 부작용과 오류를 사전에 예방하거나 신속히 시정할 수 있는 체계를 마련하고, 사용자 피드백을 기반으로 지속적으로 개선해야 한다. 또한, 명확한 불만 처리 및 분쟁 해결 절차를 제공하며, 윤리적 기준과 규제를 준수하여 시스템의 사회적 책임을 다해야 한다.

나. 실행 원칙

1) 이용자를 위한 정보 공개

이용자를 위한 정보 공개는 AI 기반 추천 알고리즘의 투명성을 실현하는 데 중요한 원칙으로, 알고리즘 작동 원리와 추천 결과의 주요 요인을 사용자에게 명확히 전달하는 것을 포함한다. 연구에 따르면, 사용자 신뢰를 높이고 알고리즘의 사회적 수용성을 강화하려면 정보의 명확성과 접근성을 보장해야 한다는 점이 강조된다(디아코푸로스, 2016). 이는 데이터 편향, 불공정 추천, 프라이버시 침해와 같은 문제를 완화하고 사용자가 시스템의 결과를 이해하고 통제할 수 있는 기반을 제공한다. 정보 공개를 보장하기 위해, 추천 알고리즘의 데이터 활용 방식과 결과 생성 매개변수를 명확히 설명하는 것이 중요하다. 실시간 알림, 팝업창, 대시보드 등을 통해 정보를 직관적으로 제공해야 하며, 데이터 활용 목적과 저장 기간을 상세히 기술한 약관과 정책이 필수적이다.

2) 이용자의 선택권 보장

이용자의 선택권 보장은 사용자가 자신의 선호와 필요에 따라 추천 시스템의 결과를 직접 통제할 수 있도록 기능을 제공하는 것을 목표로 한다. 사용자가 자신의 데이터와 추천 결과를 관리할 수 있을 때 개인화된 서비스의 만족도가 높아질 가능성이 높다. 또한, 선택

권 보장은 추천 알고리즘으로 인해 발생할 수 있는 정보 필터 버블과 편향성을 방지하는 데도 중요하다(Pariser, 2011). 이를 보장하기 위해 사용자 맞춤형 프로필 생성, 수정, 삭제 옵션과 콘텐츠 배열 및 알고리즘 선호도를 조정할 수 있는 기능이 필요하다. 이와 함께, 사용자가 자신의 시청 기록과 추천 데이터를 초기화할 수 있는 옵션을 제공해 데이터 관리의 주도권을 사용자에게 부여해야 한다.

3) 자율검증 실행

자율검증 실행은 추천 시스템의 공정성과 정확성을 유지하기 위해 지속적으로 검토하고 개선하는 과정을 포함한다. 이는 알고리즘 편향을 식별하고 완화하며, 시스템의 윤리적, 법적 준수성을 확보하기 위한 핵심 원칙으로 간주된다. 기업이 자율적으로 검증 체계를 마련하는 것은 외부 규제를 보완하고 사회적 신뢰를 구축하는 데 중요한 요소로 강조된다. 이를 보장하기 위해서는 클릭률, 유지율, 사용자 만족도와 같은 정량적 성능 평가가 정기적으로 이루어져야 한다(Jiang & Qi, 2024). 또한, 특정 그룹에 부정적 영향을 미치지 않는지 잠재적 위험을 분석하고 외부 기관의 검토와 독립적 검증을 허용하는 투명한 검증 프로세스를 마련해야 한다.

4) 불만 처리 및 분쟁 해결

불만 처리와 분쟁 해결은 사용자 경험에서 발생하는 문제를 신속히 해결하고 추천 알고리즘의 오류와 불공정성을 시정하기 위한 체계를 의미한다. 이러한 체계는 사용자 신뢰를 유지하는 데 필수적이며, 정책 보고서는 사용자 친화적인 접근 방식을 통해 문제 해결이 원활히 이루어져야 한다고 언급하고 있다. 이를 보장하기 위해 다양한 불만 제기 채널(이메일, 전화, 실시간 채팅, FAQ)을 제공해야 하며, 불만 사항을 신속히 처리할 명확한 절차와 일정이 필요하다. 또한, 24시간 고객 서비스 운영과 같은 상시 지원 체계를 통해 사용자 편의를 높여야 한다.

5) 내부 규칙 제정

내부 규칙 제정은 기업이 추천 알고리즘을 관리하기 위한 윤리적·법적 기준을 수립하고 이를 투명하게 운영하는 과정을 포함한다. 이를 보장하기 위해 데이터 보호, 편향 방지, 윤리적 설계를 포함하는 가이드라인을 마련해야 한다. 또한, 규칙의 주요 내용을 외부에 공개하고 이해관계자의 의견을 반영하며, 정기적으로 규칙을 검토하고 갱신하는 절차가

요구된다. 이러한 과정은 알고리즘의 법적, 윤리적, 안전성 측면에서 공식적인 보증을 제공하는 ‘알고리즘 감사(algorithm auditing)’의 개념과 밀접하게 연관되어 있다(Koshiyama et al., 2021).

2. 정책 및 서비스 현황 분석

정책 및 서비스 현황 분석에서는 주요 디지털 플랫폼들이 추천 알고리즘의 핵심 원칙(투명성, 공정성, 책무성)과 실행 원칙(정보 공개, 선택권 보장, 자율검증 실행, 불만 처리 및 분쟁 해결, 내부 규칙 제정)을 실제로 얼마나 잘 이행하고 있는지를 살펴볼 것이다. 이를 위해 각 플랫폼의 이용약관, 개인정보 보호정책, 내부 규정을 포함한 정책과 현재의 서비스 현황을 분석할 것이다. 구체적으로, 투명성 측면에서는 사용자가 추천 시스템의 작동 원리와 추천 결과의 이유를 쉽게 이해할 수 있도록 명확한 정보 제공과 실시간 알림이 포함되는지를 평가할 것이다. 공정성은 알고리즘 편향성을 방지하고 사용자 권익과 미디어 다양성을 보장하기 위해 어떤 기술적·관리적 조치가 이루어졌는지를 점검할 것이다. 책무성은 추천 시스템 운영 과정에서 발생하는 문제를 어떻게 해결하고 사용자와의 신뢰를 유지했는지에 초점을 맞출 것이다. 또한, 실행 원칙에서는 AI 기반 추천 시스템의 정보를 효과적으로 공개하고, 사용자에게 콘텐츠 선택권을 부여하며, 시스템의 성능과 위험성을 지속적으로 검증하는 체계가 마련되어 있는지를 평가할 것이다. 이 외에도 플랫폼이 불만 처리와 분쟁 해결을 위한 적절한 체계를 운영하고 있는지, 내부 규칙을 자율적으로 제정하고 이를 투명하게 공개했는지를 분석할 것이다. 이를 토대로 주요 플랫폼의 성과를 객관적으로 평가하며, 성공적인 사례를 발굴하고 개선이 필요한 영역을 제시하는 것을 목표로 한다.

가. 넷플릭스(Netflix)

1) 핵심원칙

(1) 투명성

넷플릭스는 인공지능 기반 추천 시스템의 투명성을 유지하기 위해 사용자에게 서비스 제공 사실을 명확히 알리고, 추천 콘텐츠의 이유를 간단히 설명하는 방식을 채택하고 있다. 예를 들어, “당신이 [이전 콘텐츠]를 보았기 때문”이라는 메시지를 통해 추천 콘텐츠가

사용자의 시청 기록과 환경 설정을 기반으로 이루어진다는 점을 명확히 전달하고 있다. 이와 같은 접근은 넷플릭스 이용 약관에서도 확인할 수 있으며, “우리는 사용자의 시청 기록과 환경 설정을 기반으로 콘텐츠를 추천합니다”라고 밝히고 있다.⁵⁾

그러나 넷플릭스는 추천 결과에 대한 불만이 제기되거나 부정적인 추천이 이루어진 경우, 해당 결과의 이유를 명확히 설명하는 체계적인 시스템을 마련하지 않고 있다. 알고리즘의 작동 원리와 추천 결과에 도달한 과정을 보다 상세히 설명하는 체계를 구축한다면, 넷플릭스는 투명성을 강화하고 사용자 신뢰를 한층 더 높일 수 있을 것이다.

(2) 공정성

넷플릭스는 추천 시스템의 공정성을 유지하기 위해 다양한 국가와 문화를 대표하는 콘텐츠를 제공하며, 사용자에게 장르와 콘텐츠에 대한 선택권을 보장하고 있다. 특히, 회사의 글로벌 콘텐츠 소개 페이지를 통해 넷플릭스가 제공하는 다양한 국가의 콘텐츠, 제작 전략, 및 글로벌 확장 계획에 대한 정보를 제공하고 있으며 여기에는 “멋진 스토리는 어디서든 만들어지고 어디서나 사랑받을 수 있기 때문에 30개 이상의 언어와 190개국에서 스트리밍되고 있습니다”라고 명시하고 있다.⁶⁾

특히, 추천 알고리즘은 딥러닝을 기반으로 사용자 선호를 반영하면서도 특정 콘텐츠나 장르에 과도하게 치우치지 않도록 설계되었으며, 이를 통해 사용자 경험의 질적 향상을 도모하고 있다. 또한, 넷플릭스는 사용자에게 추천 알고리즘의 결과를 관리할 수 있는 프로필 설정과 시청 기록 삭제 기능을 제공함으로써 추천 시스템의 공정성을 높이고 선택권을 강화하고 있다.

(3) 책무성

넷플릭스는 부모 통제 기능과 지역별 콘텐츠 큐레이션을 통해 책임 있는 서비스를 제공하고 있으며, 다양한 국가와 문화적 배경을 반영한 콘텐츠 제작으로 사회적 책무를 이행하고 있다. 사용자 불만 처리와 분쟁 해결 측면에서는 고객센터를 통해 “불법 콘텐츠나 저작권 침해 신고를 신속히 처리한다”는 정책을 운영하고 있으며, 사용자 요청에 대한 체계

5) 넷플릭스, <https://help.netflix.com/legal/termsofuse>

6) 넷플릭스, <https://about.netflix.com/ko>

적인 대응 체계를 마련하고 있다.⁷⁾

또한, 기능적 오류가 발생할 경우 사용자에게 직접 알림을 제공하거나 시스템 업데이트를 통해 문제를 신속히 시정함으로써 안정적이고 책임 있는 플랫폼 운영을 보장하고 있다. 이러한 노력으로 넷플릭스는 사회적 책무, 분쟁 해결, 오류 시정에서 높은 평가를 받고 있으나, 알고리즘 관련 정보를 사용자와 더 투명하게 소통함으로써 이용자 신뢰를 한층 더 강화할 필요가 있다.

2) 실행원칙

(1) 이용자를 위한 정보 공개

넷플릭스는 개인화 추천 시스템 사용 사실을 명확히 명시하고 있으며, 시청 기록과 선호도에 기반한 추천이라는 점을 약관에서 설명하고 있다. 예를 들어, “우리는 사용자의 시청 기록과 환경 설정을 기반으로 콘텐츠를 추천합니다”라는 문구를 통해 사용자가 추천 서비스의 원리를 이해할 수 있도록 하고 있다.⁸⁾

그러나 AI 사용 사실이나 추천 시스템 작동 방식에 대한 정보를 실시간 알림이나 팝업창과 같은 사용자 친화적인 방식으로 통지하지는 않고 있다. 추천 알고리즘에 영향을 미치는 배열 요소 역시 약관을 통해 일부 설명하고 있으나, 흡화면 알림이나 팝업창으로 명확히 고지하지 않아 접근성이 부족한 상태다. 넷플릭스는 AI 기반 추천 시스템 사용 사실을 명확히 알리고 있지만, 실시간 알림 및 팝업창 형태의 정보 제공을 강화함으로써 사용자 이해와 접근성을 높일 필요가 있다.

(2) 이용자의 선택권 보장

넷플릭스는 사용자가 자신의 선호 콘텐츠를 직접 선택하고 관리할 수 있는 선택권을 보장하고 있으며, 이를 위해 프로필 설정 기능과 시청 기록 삭제 옵션을 제공하고 있다. 이러한 접근은 사용자가 추천 시스템의 결과를 어느 정도 제어할 수 있도록 지원하며, 개인화된 시청 경험을 강화하는 데 기여하고 있다. 그러나 콘텐츠 배열 방식에 영향을 미치는 주요 요소를 사용자가 직접 수정하거나 변경할 수 있는 기능은 부족한 상황이다. 콘텐츠

7) 넷플릭스, <https://help.netflix.com/ko>

8) 넷플릭스, <https://help.netflix.com/legal/termsofuse>

추천 이유를 설명하는 기능은 제공되지만, 배열 방식을 조정할 수 있는 사용자 친화적인 기능이 부재한 부분은 개선이 필요해 보인다.

(3) 자율검증 실행

넷플릭스는 추천 시스템의 자율검증 체계를 강화하기 위해 위험성을 사전에 예측하고 평가하며, 알고리즘이 사용자 선호를 반영하면서도 특정 콘텐츠에 편향되지 않도록 설계하고 있다. “회원이 넷플릭스 서비스에 액세스할 때마다 의견을 수집하여 해당 신호로 넷플릭스 알고리즘을 계속 개선함으로써, 회원 시청 가능성이 가장 높은 콘텐츠를 점점 더 정확하게 예측합니다”라는 설명처럼, 넷플릭스는 알고리즘의 성능과 잠재적 위험성을 지속적으로 점검하고 있다.⁹⁾

추천 결과의 유효성과 사용자 만족도를 정기적으로 확인하며, 데이터 활용 방식에 대한 내부 평가를 시행해 추천 알고리즘의 품질을 관리하고 있다. 또한, 개발 및 테스트 과정을 내부적으로 기록하고 관리하지만, 외부 검증 가능성을 높이기 위한 세부적인 정보 공개는 부족한 상황이다. 종합적으로 넷플릭스는 자율검증 체계를 통해 위험 평가와 지속적 관리를 효과적으로 수행하고 있으나, 외부 검증 가능성과 접근성을 강화하여 투명성과 신뢰성을 더욱 높일 필요가 있다.

(4) 불만 처리 및 분쟁 해결

넷플릭스는 고객 불만 처리와 분쟁 해결에서 높은 수준의 신속성과 접근성을 제공하는 체계를 운영하고 있다. 고객이 언제든지 불편사항을 보고하고 지원받을 수 있도록 다양한 지원 채널을 마련했으며, 고객센터는 한국어 상담의 경우 매일 오전 8시부터 오후 8시까지, 영어 상담은 연중무휴 24시간 운영된다. 또한, 고객은 전화 상담과 실시간 채팅 서비스를 통해 문제를 해결할 수 있으며, 웹사이트나 모바일 앱을 통해 콘텐츠 관련 문제를 직접 신고할 수 있다. 이러한 채널을 통해 버퍼링, 자막, 음성, 화질 등 다양한 문제를 신속히 처리하고 있으며, 무료 상담 시스템을 운영해 고객 편의를 높이고 있다. 그러나, 홈 화면 알림이나 팝업창과 같은 실시간 접근 안내를 추가하여 접근성을 더욱 강화할 수 있을 것으로 보인다.

9) 넷플릭스, <https://help.netflix.com/ko/node/100639>

(5) 내부 규칙 제정

넷플릭스는 데이터 보호와 추천 알고리즘 관리에 관한 내부 규칙을 일부 공개하며, 알고리즘 편향성을 방지하기 위한 가이드라인을 운영하고 있다. 특히, 넷플릭스는 개인정보 처리방침을 통해 사용자 데이터의 수집 및 사용 목적을 명확히 설명하고 있으며, 구체적으로 개인화된 콘텐츠 추천 제공, 서비스 개선 및 최적화, 계정 관리 및 운영, 연구 및 분석, 안전과 보안 강화, 사기 방지, 법적 요구사항 준수 등을 목적으로 데이터를 수집 및 사용한다고 명시하고 있다.¹⁰⁾

수집되는 데이터 유형에는 개인 정보, 결제 정보, 계정 및 프로필 정보, 사용 정보, 광고 정보, 기기 및 네트워크 정보, 통신 내용 등이 포함되며, 이러한 데이터를 통해 사용자 맞춤형 서비스를 제공하고 서비스 품질을 유지·개선하고 있고 밝히고 있다. 따라서, 넷플릭스는 데이터 활용에 대한 명확한 목적을 설정하고 이를 방침에 구체적으로 기술하여 사용자 신뢰를 구축하고 있는 것으로 보인다.

나. 디즈니 플러스(Disney+)

1) 핵심원칙

(1) 투명성

디즈니 플러스는 추천 알고리즘이 사용자의 시청 기록과 선호도를 기반으로 작동한다고 명시하며, 사용자에게 데이터 활용 방식을 이해할 수 있는 정보를 제공하고 있다. 예를 들어, 디즈니+ 개인정보 보호정책에서 “귀하의 시청 기록과 선호도는 맞춤형 콘텐츠 추천에 사용됩니다”라는 내용을 통해 데이터 활용 원칙을 투명하게 전달하고 있다.¹¹⁾

또한, 추천 결과에 대한 이용자 불만이나 부정적인 요소가 발생했을 때 이를 해결하기 위한 시스템을 운영함으로써 문제 해결에 주력하고 있다. 그러나 추천 과정의 상세한 설명을 사용자에게 제공하는 데에는 한계가 있으며, 알고리즘이 추천 결과에 도달하는 방식에 대한 구체적인 정보는 부족하다. 종합적으로 디즈니+는 추천 시스템의 기본 작동 원리를 명확히 공개하여 투명성을 유지하고 있으나, 추천 결과의 도달 과정에 대한 상세한 설명을 강화함으로써 사용자 신뢰를 더욱 높일 필요가 있어 보인다.

10) 넷플릭스, <https://help.netflix.com/legal/privacy>

11) 디즈니플러스, <https://www.disneyplus.com/legal/privacy-policy>

(2) 공정성

디즈니 플러스는 가족 친화적 콘텐츠와 연령 제한 설정을 통해 다양한 사용자 그룹의 권익을 보호하며, 모든 연령층이 이용할 수 있는 콘텐츠 제공을 핵심 목표로 설정하고 있으며, 이는 디즈니+ 콘텐츠 정책의 문구에 명확히 밝히고 있다.¹²⁾

또한, 디즈니+는 알고리즘의 사전적·사후적 평가를 통해 특정 연령대나 취향에 과도하게 치우치지 않는 콘텐츠 큐레이션을 목표로 하고 있으며, 이를 통해 시스템의 공정성을 지속적으로 개선하고 있다. 사용자 선택권 보장 측면에서도 디즈니+는 프로필별 콘텐츠 접근 제한과 개인화 설정을 통해 사용자가 선호도와 콘텐츠 접근성을 직접 관리할 수 있도록 지원하고 있다. 종합적으로 디즈니+는 가족 중심의 접근 방식과 다양한 연령층을 고려한 콘텐츠 제공으로 공정성을 확보하고 있으며, 알고리즘 평가와 지속적 개선에서 강점을 보이는 플랫폼으로 평가된다.

(3) 책무성

디즈니 플러스는 부모 통제 기능과 콘텐츠 연령 제한 설정을 통해 부적절한 콘텐츠 접근을 방지하고, 가족 친화적 정책을 강조함으로써 사회적 책무를 다하고 있다. 이러한 접근은 디즈니+가 모든 연령층의 사용자를 고려한 플랫폼 운영에 주력하고 있음을 보여준다. 또한, 디즈니+는 전화 상담, 실시간 온라인 채팅, 헬프 센터 등 다양한 고객 지원 방법을 통해 신고된 문제를 처리하고 있다.¹³⁾

기능적 오류 관리 측면에서도 디즈니+는 정기적인 시스템 점검과 업데이트를 통해 오류 발생을 최소화하고, 기술적 문제가 발생했을 때 사용자에게 해결 방법을 명확히 안내하고 있다. 종합적으로 디즈니+는 가족 중심의 사회적 책임을 다하기 위해 체계적인 정책과 기술적 대응 시스템을 운영하며, 분쟁 해결과 기술적 오류 관리에서 높은 신뢰도를 보여주는 플랫폼으로 평가된다.

2) 실행원칙

(1) 이용자를 위한 정보 공개

디즈니 플러스는 개인정보 보호정책에서 AI 기반 개인화 추천 시스템의 존재를 명확히

12) 디즈니플러스, <https://www.disneyplus.com/legal/privacy-policy>

13) 디즈니플러스, <https://help.disneyplus.com>

설명하며, 사용자가 추천 시스템의 작동 원리를 이해할 수 있도록 정보를 제공하고 있다.¹⁴⁾

구체적으로 디즈니는 사용자가 보는 콘텐츠, 서비스 이용 빈도, 선호도 등 사이트와 애플리케이션에서의 활동 정보를 수집하며, 이러한 데이터를 기반으로 콘텐츠와 경험을 개인화하고 있음을 명시하고 있다. 콘텐츠 배열의 주요 요인에 대해서도 약관이나 개인정보 보호 정책에서 간략하게나마 설명하고 있으나, 이에 대한 구체적인 정보 제공은 부족해 보인다.

(2) 이용자의 선택권 보장

디즈니 플러스는 사용자 프로필 설정을 통해 선호도 기반 추천을 설정할 수 있도록 하며, 연령 제한과 콘텐츠 접근 권한을 관리할 수 있는 기능을 제공하여 선택권을 보장하고 있다. 특히, 디즈니+는 부모 통제 기능을 통해 사용자가 자녀의 콘텐츠 접근을 효과적으로 관리할 수 있도록 다양한 도구를 제공하고 있다. 주요 기능으로는 키즈 프로필이 있어, 연령에 적합한 콘텐츠만을 제공하는 어린이용 프로필을 생성할 수 있으며, 이를 통해 부적절한 콘텐츠 접근을 차단하고 안전한 시청 환경을 조성할 수 있다. 또한, 콘텐츠 등급 제한 기능을 통해 0+부터 18+까지의 다양한 연령 등급을 설정하여 특정 연령층에 부적합한 콘텐츠를 필터링할 수 있다.¹⁵⁾

그러나 콘텐츠 추천의 주요 요소를 정책에서 명시하고 있음에도 불구하고, 사용자 스스로 콘텐츠 배열 요소를 직접 변경할 수 있는 기능은 제공되지 않고 있다. 종합적으로 디즈니+는 연령 및 선호도 관리 기능을 통해 사용자 선택권 보장에서 강점을 보이고 있으나, 콘텐츠 배열 변경 기능을 제공하지 않는 점은 한계로 지적되며, 이 부분에서 추가적인 개선이 필요하다.

(3) 자율검증 실행

디즈니 플러스는 자율검증 실행을 통해 콘텐츠 추천 과정에서의 위험성을 사전에 예측하고 평가하며, 사용자 안전을 우선시하고 있다. 연령 제한과 가족 중심 콘텐츠 정책을 활용하여 부적절한 콘텐츠 접근을 방지하는 디즈니+ 콘텐츠 정책을 통해 알고리즘 설계의 기본원칙을 명시하고 있다. 그러나, 외부 검증 가능성을 위한 정보 공개는 부족한 상태다. 종합적으로 디즈니+는 위험 관리와 지속적 검토에서 높은 수준을 유지하고 있으며, 안정

14) 디즈니플러스, <https://www.disneyplus.com/legal/privacy-policy>

15) 디즈니플러스, <https://www.astrosafe.co/blog/disney-plus-parental-controls>

적인 자율검증 체계를 갖추고 있지만, 자율검증 과정의 외부 투명성을 강화하여 사용자 신뢰를 더욱 높일 필요가 있다.

(4) 불만처리 및 분쟁 해결

디즈니 플러스는 고객 지원 센터를 통해 사용자 불만 처리와 문제 해결을 위해 체계적이고 다양한 고객 지원 시스템을 운영하고 있다. 헬프 센터는 자주 묻는 질문(FAQ)과 문제 해결 가이드를 통해 사용자가 간단한 문제를 스스로 해결할 수 있도록 돕는다. 또한, 실시간 채팅 기능을 통해 고객 지원 담당자와 직접 대화할 수 있는 기회를 제공하며, 사용자와의 실시간 소통을 강화하고 있다. 24시간 운영되는 전화 상담 서비스는 시간에 구애받지 않고 문제를 신속히 해결할 수 있도록 지원한다. 이러한 다각적인 지원 체계는 디즈니+가 사용자 불만 처리의 신속성과 효율성을 중시하고 있음을 보여준다. 다만, 홈화면 알림이나 팝업창을 통한 실시간 접근 안내와 같은 사용자 친화적 추가 기능을 도입한다면, 고객 지원 시스템의 접근성이 더욱 강화될 것으로 보인다.

(5) 내부 규칙 제정

디즈니 플러스는 데이터 보호 및 추천 서비스 운영에 관한 내부 규칙을 일부 공개하며, 사용자 보호를 강화하기 위한 연령 제한과 관련 정책을 명확히 설정하고 있다. 또한, 디즈니의 모회사는 개인정보를 “주의와 존중”으로 다루겠다는 원칙을 명시하며, 사용자의 신뢰를 유지하기 위해 노력하고 있다. 개인정보 처리방침에는 수집되는 정보의 유형, 정보 수집 방법, 정보 사용 방식, 제한적인 정보 공유 상황, 사용자에게 제공되는 통제권과 선택권, 데이터 보안과 무결성, 데이터 보존 정책 및 글로벌 데이터 전송·저장·처리 관행 등이 포함되어 있다. 이러한 조치는 개인정보와 서비스 데이터를 안전하게 관리하기 위한 디즈니의 체계적인 접근 방식을 보여준다. 비록 정책에서 구체적인 문구를 활용하지는 않았지만, 디즈니는 개인정보 보호의 중요성을 인식하고 이를 실현하기 위해 명확하고 종합적인 보호 체계를 구축하고 있음을 알 수 있다.¹⁶⁾

16) 월트디즈니컴퍼니, <https://privacy.thewaltdisneycompany.com/en/>

다. 웨이브(Wavve)

1) 핵심원칙

(1) 투명성

웨이브는 개인정보 처리방침을 통해 사용자 데이터를 활용한 맞춤형 서비스 제공에 대해 상세히 설명하고 있다.¹⁷⁾

서비스 운영 목적으로 “맞춤형 정보추천”을 위해 이메일 주소, 기기ID(모바일 단말기), 시청 기록 등을 수집 및 처리하며, 이는 사용자의 관심사에 기반한 맞춤형 콘텐츠 정보와 광고 서비스를 제공하기 위함이다. 행태정보 수집 및 이용에 관한 안내에서는 “회원 관심 기반 맞춤형 콘텐츠 정보추천 서비스”와 “회원 관심 기반 맞춤형 콘텐츠 광고 서비스”를 목적으로 기기ID와 시청 기록을 수집한다고 명시하고 있다. 수집된 정보는 회원 탈퇴 후 3개월(정보 추천 서비스) 또는 회원 탈퇴 및 동의 철회 시까지(광고 서비스) 보유 및 이용되며, 웨이브는 사용자에게 브라우저나 앱 설정을 통해 이러한 정보 수집을 거부할 수 있는 통제권을 제공한다. 이러한 방침은 개인정보 처리와 관련한 웨이브의 투명성을 보여주며, 사용자가 데이터 활용 방식에 대해 명확히 이해하고 제어할 수 있도록 설계되어 있다. 그러나 추천 과정의 상세한 설명과 더불어 사용자 불만을 체계적으로 처리할 수 있는 시스템이 마련된다면, 플랫폼 신뢰성과 사용자 경험이 더욱 향상될 것으로 보인다.

(2) 공정성

웨이브는 다양한 장르와 콘텐츠를 제공하며, 사용자 데이터를 활용해 균형 잡힌 콘텐츠 큐레이션을 통해 편향성을 방지하고 있다. “웨이브는 알고리즘 딥러닝 기술을 활용해 사용자 맞춤형 추천 시스템을 강화하고 있다”라는 연합뉴스의 보도처럼, 웨이브는 딥러닝 기술을 통해 개인화된 추천 서비스를 최적화하면서도 콘텐츠 다양성을 유지하려는 노력을 기울이고 있다.¹⁸⁾

또한, 알고리즘의 정기적인 점검과 개선을 통해 데이터 편향성을 방지하고, 추천 시스템의 공정성을 확보하기 위해 기술적·관리적 조치를 지속적으로 시행하고 있다. 사용자는 프로필 설정 기능을 활용해 개인화된 콘텐츠 선호를 직접 관리할 수 있어, 콘텐츠 접근 방

17) 웨이브, <https://member.wavve.com/signup/terms>

18) 연합뉴스(2020. 9. 28.), “토종OTT 웨이브 회원 1천 만 돌파… ‘추천 기능 최적화’”

식에 대한 선택권도 보장받는다. 종합적으로, 웨이브는 추천 알고리즘의 공정성을 유지하고자 하는 노력을 통해 편향성을 방지하고 있으며, 기술적 조치와 사용자 선택권 보장 측면에서 강점을 보이는 플랫폼으로 평가된다.

(3) 책무성

웨이브는 부모 통제 기능과 데이터 보호 정책을 통해 책임 있는 플랫폼 운영을 강조하며, 지역별 콘텐츠 큐레이션을 통해 사용자 신뢰를 구축하고 있다. 이러한 사회적 책무는 가족 친화적 접근과 데이터 관리의 투명성을 통해 실현되고 있다. 또한, 웨이브는 기술적 오류 시정과 분쟁 해결에서 체계적이고 효과적인 시스템을 갖추고 있으며, 책임 있는 운영 방식을 통해 사회적 책무를 충실히 이행하고 있는 플랫폼으로 평가된다. 다만, 오류 관리와 소통 체계의 접근성을 더욱 강화한다면 사용자 경험이 한층 향상될 수 있을 것이다.

2) 실행원칙

(1) 이용자를 위한 정보 공개

앞서 기술한 대로 웨이브는 개인정보 처리방침을 통해 사용자 데이터를 활용한 맞춤형 서비스 제공에 대해 상세히 설명하고 있다. 사용자에게 이러한 정보가 어떻게 추천 시스템에 반영되고, 추천 결과를 도출하는지에 대한 상세한 정보를 실시간으로 제공하지는 않는다. 이러한 점에서 웨이브는 데이터 활용의 기본적인 원칙을 공개하고 있으나, 사용자 친화적 정보 전달 방식과 추천 알고리즘 작동 원리에 대한 더 구체적인 설명이 보완될 필요가 있다. 이를 통해 사용자 신뢰를 높이고 정보 공개의 투명성을 강화할 수 있을 것이다.

(2) 이용자의 선택권 보장

웨이브는 이용자의 선택권 보장을 위해 다양한 보호 체계와 정책을 수립하여 운영하고 있다. 웨이브는 이용자 보호 및 서비스 품질 향상을 위한 비전을 수립하고, 이용자 피해 예방을 목표로 보호 체계를 마련하며, 정당한 의견이나 불만을 적극적으로 처리하고 있다.¹⁹⁾

선택권 보장을 위한 구체적인 계획으로는 이용자 보호 책임자와 담당자 지정을 통해 사업 계획 수립과 개선 방안을 마련하며, 관련 법률과 정부 제도를 준수해 이용자 보호 정책을 실행한다. 정기 교육을 통해 서비스 리스크를 점검하고, 신규 서비스 출시 전에 리스크

19) 웨이브, <https://www.contentwavve.com/news/articleView.html?idxno=587>

체크와 이용자 보호 인식을 제고하며, 운영 상황을 상시 점검하고 개선하는 프로세스를 구축하고 있다. 또한, 유해 정보와 환경으로부터 이용자를 보호하기 위해 보안 솔루션과 위협 정보 분석 체계를 도입하며, 이용자의 프라이버시를 보호하기 위한 기술적 조치를 마련하고 있다. 이러한 노력은 웨이브가 이용자 권리 보호와 피해 예방을 핵심 가치로 삼고, 이용자에게 선택권과 통제권을 보장하기 위한 실질적인 접근 방식을 취하고 있음을 보여준다.

(3) 자율검증 실행

웨이브는 추천 알고리즘의 성능을 평가하고 개선하기 위해 체계적인 접근 방식을 채택하고 있다. 웨이브의 내부 연구원의 인터뷰에 따르면, 웨이브는 테스트를 통해 알고리즘 성능을 정기적으로 평가하며, 사용자를 랜덤으로 분류해 기존 추천 밴드와 개선된 추천 밴드를 각각 제공하고, 클릭률, 시청률, 추천 콘텐츠의 다양성 등의 지표를 기반으로 사용자 반응을 비교한다고 한다.²⁰⁾

또한, 웨이브는 사용자 시청 시간(북마크 데이터), 활동 기록(로그 데이터), 콘텐츠 정보(메타 데이터) 등 다양한 데이터를 활용해 개인화된 추천 시스템을 지속적으로 개발하고 있고 밝혔다. 이러한 접근은 알고리즘 성능 향상과 사용자 경험 개선을 목표로 하며, 데이터를 활용한 정량적 평가와 안전성을 고려한 테스트 과정을 통해 안정적이고 신뢰할 수 있는 추천 시스템을 구축하려는 웨이브의 체계적 자율검증 노력을 보여준다.

(4) 불만 처리 및 분쟁 해결

웨이브는 불만 처리 및 분쟁 해결 측면에서 이용자 중심의 체계를 운영하며, 신속하고 명확한 대응을 통해 사용자 신뢰를 구축하고 있다. 분쟁 해결과 관련해, 웨이브는 개인정보분쟁조정위원회와 한국인터넷진흥원 개인정보침해신고센터 등의 외부 기관과 협력하여 개인정보 침해 문제를 해결할 수 있도록 지원한다.²¹⁾ 이러한 접근은 이용자가 적시에 구제를 받을 수 있는 체계를 보장하며, 사용자 권리를 보호하기 위한 웨이브의 의지를 반영한다.

웨이브는 또한 사이버 감사실을 운영하여 윤리 경영을 실천하며, 명확한 제보 처리 절차

20) 웨이브, <https://wavve-comm.tistory.com/139>

21) 웨이브, <https://member.wavve.com/signup/terms>

차를 통해 투명성을 확보하고 있다.²²⁾ 이를 통해 이용자 불만 사항이나 신고가 체계적으로 관리되며, 서비스 품질 향상을 위한 피드백으로 활용된다. 종합적으로, 웨이브는 신고된 문제의 신속한 처리와 사용자와의 소통을 기반으로 한 분쟁 해결 체계를 갖추고 있으며, 이용자 권리 보호와 피해 예방을 위해 기술적·관리적 조치를 지속적으로 강화하고 있다. 이러한 노력은 웨이브가 신뢰와 책임을 중시하는 플랫폼임을 보여준다.

(5) 내부 규칙 제정

웨이브는 개인정보 처리와 데이터 보호를 위한 내부 규칙을 운영하며, 개인정보 처리방침을 통해 이러한 운영 방침을 일부 명시하고 있다. 데이터 보호와 알고리즘 개선을 위한 기술적·관리적 절차를 유지하며, 서비스 편향성을 관리하기 위해 체계적인 노력을 기울이고 있다. 또한, 내부 검증 시스템과 고객센터를 통해 사용자 문제를 해결하고, 분쟁 해결 프로세스를 지원하고 있다. 그러나 내부 규칙과 자율검증 체계에 관한 세부 내용은 외부에 명확히 공개되지 않으며, 관련 조직이나 체계에 대한 구체적인 정보도 제한적이다. 이는 사용자와 외부 이해 관계자들에게 플랫폼의 투명성과 신뢰성을 높이는 데 장애 요인이 될 수 있다.

라. 티빙(Tving)

1) 핵심원칙

(1) 투명성

티빙은 개인화된 콘텐츠 추천 서비스를 제공하며, 데이터 활용의 투명성을 높이기 위한 다양한 노력을 기울이고 있다. 특히, 티빙은 콘텐츠 추천 서비스 고도화 작업을 통해 개인화 서비스를 강화하고, 콘텐츠 사업자(CP)와의 협력을 강조한 ‘티빙 2.0 스케일업’ 전략을 통해 플랫폼의 개방성을 확대했다.²³⁾ 이를 통해 데이터 활용과 추천 알고리즘 운영 방식에서 투명성을 확보하려는 의지를 드러냈다. 예를 들어, 사용자가 선호하는 콘텐츠를 직접 선택하도록 하여 콘텐츠 추천 과정에 사용자 참여를 유도하고, 추천 알고리즘이 사용자 데이터를 어떻게 활용하는지 간접적으로 경험할 수 있게 했다. 그러나 추천 알고리즘의

22) 웨이브, <https://www.contentwavve.com/com/com-5.html>

23) 시사저널e(2020. 11. 11.), “OTT업계, 콘텐츠 추천 정확도 높이기 총력”

작동 원리와 데이터 처리 과정에 대한 구체적이고 사용자 친화적인 설명은 여전히 부족한 상태로, 이를 보완한다면 티빙은 투명성을 더욱 강화하며 사용자 신뢰를 높일 수 있을 것이다.

(2) 공정성

티빙은 다양한 장르의 콘텐츠를 제공하며, 특정 사용자 그룹이나 취향에 치우치지 않는 공정한 큐레이션을 목표로 하고 있다. 사용자의 선호도를 반영한 맞춤형 광고 도입을 검토 중이며, 이를 통해 사용자 경험을 향상시키고 있다는 서울경제의 보도는 티빙이 사용자 데이터를 기반으로 콘텐츠와 광고의 개인화를 지속적으로 개선하고 있음을 보여준다.²⁴⁾ 티빙은 알고리즘의 사전적·사후적 영향 평가를 통해 데이터 편향성을 방지하고, 시스템 신뢰도를 유지하며, 지속적인 알고리즘 개선을 통해 공정성을 확보하기 위한 노력을 기울이고 있다. 또한, 사용자는 프로필 설정을 통해 콘텐츠 접근과 추천 알고리즘 결과를 관리할 수 있어 개인화 추천과 공정성 사이에서 균형을 유지하는 데 기여하고 있는 것으로 보인다.

(3) 책무성

티빙은 부모 통제 및 콘텐츠 접근 제한 기능을 통해 가족 중심의 안전한 사용 환경을 조성하며, 사회적 책무를 충실히 이행하고 있다. 프로필 잠금 설정 기능을 통해 부모는 자녀의 콘텐츠 접근을 제한할 수 있으며, 청소년 관람 불가 콘텐츠는 성인 인증을 완료한 프로필에서만 접근 가능하다. 또한, 성인 인증이 이루어지지 않은 프로필을 자녀용으로 설정하면, 자녀가 부적절한 콘텐츠에 노출되지 않도록 보호할 수 있다. 프로필 잠금 설정은 개인 맞춤형 콘텐츠 추천에도 기여하며, 어린 자녀가 불필요한 콘텐츠에 노출되지 않도록 효과적으로 방지한다. 이러한 기능은 부모가 자녀의 콘텐츠 소비를 간접적으로 통제할 수 있도록 돕는 동시에, 사용자 신뢰를 구축하는 데 중요한 역할을 한다. 티빙은 이러한 기능을 통해 안전한 콘텐츠 환경을 제공하며, 플랫폼이 사용자 보호와 신뢰 구축을 위해 지속적으로 노력하고 있음을 보여준다.

24) 서울경제(2023. 5. 17.), “티빙, 맞춤형 광고 도입 검토…수익성 확 끌어올린다”

2) 실행원칙

(1) 이용자를 위한 정보 공개

티빙은 AI 기반 추천 시스템 사용 사실을 명시적으로 공개하고 있으나, 추천 알고리즘의 주요 요소를 실시간으로 제공하거나 사용자 친화적인 접근성을 강화하는 데는 개선이 필요하다. 이러한 보완을 통해 플랫폼의 투명성과 사용자 신뢰를 한층 강화할 수 있을 것이다.

(2) 이용자의 선택권 보장

티빙은 사용자 프로필 설정을 통해 콘텐츠 선호를 관리할 수 있는 기능을 제공하며, 데이터 초기화와 개인화 설정을 지원함으로써 사용자 선택권 보장에 기여하고 있다. 회원은 언제든지 추천 콘텐츠 설정을 변경하거나, 데이터를 초기화할 수 있다는 개인정보 처리방침의 문구는 사용자가 자신의 데이터와 추천 알고리즘 설정을 적극적으로 관리할 수 있음을 보여준다.²⁵⁾ 그러나 콘텐츠 배열 방식에 대한 직접적인 변경 옵션은 제한적이며, 추천 알고리즘의 주요 요소를 사용자가 수정하거나 조정할 수 있는 기능은 부족한 상태다. 이는 사용자가 콘텐츠 추천 방식에 대한 더 큰 통제권을 갖기 위해 개선이 필요한 부분으로 지적된다.

(3) 자율검증 실행

티빙은 사용자 데이터를 활용해 플랫폼과 서비스를 지속적으로 개선하고 있다. 회사의 한 관계자는 “고객 불평을 해결할 뿐만 아니라, 고객의 요구를 미리 파악하고 잠재적 문제를 예측해 개발에 반영하겠다”고 언급하며, 사용자 피드백과 데이터 분석을 기반으로 문제를 해결하고 플랫폼 기능을 최적화할 의지를 밝혔다.²⁶⁾ 티빙은 이러한 기술적 접근을 통해 사용자 경험을 개인화하고 개선하며, 추천 서비스의 정확성과 효과를 강화하고 있다. 이와 같이 티빙은 자율검증 실행에서 적절한 체계를 운영하며, 추천 서비스의 적절성을 평가하고 관리하는 데 성과를 보이고 있는 것으로 평가된다.

(4) 불만 처리 및 분쟁 해결

티빙은 고객 불만 처리와 분쟁 해결을 위해 다양한 채널을 운영하며, 사용자 편의를 우

25) 티빙, <https://www.tving.com/policy/terms>

26) 아이뉴스24(2022. 11. 24.), “접속은 신속, 화질은 향상, 추천은 똑똑’… 티빙, UX 경험 확 바꾼다.”

선시하고 있다. 전화 상담, 카카오톡 실시간 채팅 상담, 이메일 문의, 웹사이트 및 모바일 앱의 1:1 문의와 FAQ 확인 등의 채널을 통해 고객의 다양한 요구를 수용하고 있다. 또한, 고객의 정당한 의견과 불만을 적극적으로 처리하고, 즉시 해결이 어려운 경우 사유와 처리 일정을 명확히 안내하는 체계를 마련했다. 청약철회의 경우, 이용권 구매 후 7일 이내에 사용하지 않은 경우 전액 환불을 보장하여 고객의 권익을 보호하고 있다. 이러한 노력은 사용자 신뢰를 높이고, 문제 해결 과정에서 투명성과 접근성을 확보하려는 티빙의 노력을 보여준다. 그러나 실시간 알림 시스템과 같은 추가적인 접근성 강화 조치는 사용자 경험을 더욱 향상시킬 여지를 남긴다.

(5) 내부 규칙 제정

티빙은 추천 알고리즘 개발과 운영에서 개인화 추천을 중심으로 데이터 수집, 분석, 그리고 모델 개선에 중점을 두고 있다. 알고리즘 개발을 담당하는 회사 관계자와의 인터뷰에 따르면 티빙은 고객의 취향 정보를 수집하고 기존 사용 패턴을 분석하여 콜라보레이션 및 딥러닝 기반 추천 모델을 연구하고 적용하며, 추천 모델의 효과를 측정하고 개인 맞춤형 피팅을 수행하는 전체 파이프라인을 구축하여, 사용자 성향을 분석하고, 이를 바탕으로 데이터를 정리, 저장, 분석하여 사용자 피드백을 반영해 추천 모델을 개선하는 에코 시스템을 운영하고 있다고 한다.²⁷⁾ 그러나 이러한 노력에도 불구하고, 추천 알고리즘과 관련된 내부 규칙이나 명확한 정책에 대한 구체적인 정보는 공개되지 않았다. 이는 사용자와 외부 이해 관계자들에게 알고리즘의 공정성과 투명성을 증명하기 위한 추가적인 조치가 필요함을 시사한다.

마. 페이스북(Facebook) 및 인스타그램(Instagram)

1) 핵심원칙

(1) 투명성

페이스북과 인스타그램은 추천 서비스의 투명성을 유지하기 위해 사용자에게 주요 정보를 명확히 공개하고, 이를 지원하기 위한 체계적인 접근 방식을 채택하고 있다. 대표적으로 “이 광고가 표시되는 이유는?(Why am I seeing this ad?)”과 “이 제시물이 표시되는 이유

27) 테크수다(2021. 10. 14), “조성철 CTO가 전하는 티빙 고객 취향 파악하기 시즌1”

는?(Why am I seeing this post?)” 기능은 사용자가 콘텐츠와 광고 추천이 이루어진 주요 요인을 이해할 수 있도록 돕는 기능으로, 사용자가 광고 타겟팅 및 콘텐츠 노출의 이유를 쉽게 확인할 수 있도록 설계되었다.²⁸⁾ 이 기능을 통해 사용자는 광고주가 자신을 타겟팅한 이유, 사용된 관심사나 카테고리 정보, 타겟팅 정보의 출처(예: 방문한 웹사이트, 다운로드한 앱, 좋아요 누른 페이지 등)를 상세히 확인할 수 있다. 또한, 사용자의 기본 인구통계 정보(예: 나이, 성별, 위치 등)가 타겟팅에 어떻게 사용되었는지와 광고주가 사용자의 개인 정보를 업로드했는지 여부를 확인할 수 있다. 더 나아가, 광고주가 다른 마케팅 파트너와 협력했는지에 대한 정보도 제공된다. 이러한 투명성 기능은 사용자가 광고 타겟팅 과정을 이해하고 자신의 광고 경험을 조정할 수 있는 선택권을 부여하며, 광고 시스템의 투명성을 높이고 사용자 신뢰를 강화하려는 페이스북과 인스타그램의 노력의 일환이다. 이를 통해 플랫폼은 광고와 콘텐츠 노출에 대한 사용자의 신뢰를 지속적으로 구축하고 있다.

(2) 공정성

페이스북과 인스타그램은 커뮤니티 규정을 통해 차별적 콘텐츠와 편향성을 방지하며, 다양한 관점을 반영하는 콘텐츠를 노출함으로써 공정성을 유지하고 있다.²⁹⁾ 특히, 폭력적이거나 범죄 행위를 조장하거나 미화하는 콘텐츠, 증오심 표현, 그리고 괴롭힘 및 따돌림을 엄격히 금지하며, 이를 모니터링하고 제거하는 체계를 운영하고 있다. 폭력 및 범죄 행위와 관련하여, 자살, 자해, 동물 학대 등을 조장하거나 묘사하는 게시물은 금지 대상으로 명확히 규정하고 있으며, 이를 방지하기 위한 모니터링을 강화하고 있다. 증오심 표현의 경우, 인종, 종교, 성적 지향, 성별 정체성 등 보호받는 특성을 기반으로 한 직접적인 공격을 금지하며, 증오심 표현을 정의하고 허용하지 않는 명확한 방침을 세우고 있다. 괴롭힘 및 따돌림에 관해서도 어떠한 형태의 사이버 괴롭힘도 용납하지 않으며, 관련 콘텐츠를 제거하거나 제한하는 조치를 시행하고 있다. 사용자들에게는 부적절한 콘텐츠를 발견했을 때 이를 신고하도록 권장하고, 신고 절차를 간소화하여 게시물의 점(...) 메뉴를 통해 ‘지원 찾기 또는 게시물 신고’를 선택해 신고할 수 있는 기능을 제공한다. 이러한 접근은 사용자

28) 메타, <https://about.fb.com/ko/news/2019/06/facebook이-어떻게-작동하는지-보다-나은-이해를-돕기-위해/>

29) 메타, <https://transparency.meta.com/policies/community-standards/>

와 플랫폼 간의 협력을 통해 규정 준수를 강화하려는 노력의 일환으로, 공정성을 유지하고 커뮤니티의 안전과 다양성을 보장하려는 페이스북의 방침을 보여준다.

(3) 책무성

페이스북과 인스타그램은 사회적 책무 또한 위 커뮤니티 규정을 통해 이행하고 있는 것으로 평가된다. 구체적으로 페이스북은 콘텐츠 검토 시스템을 통해 사용자 신고, AI 기술, 인간 검토를 결합하여 부적절한 콘텐츠를 효과적으로 관리하고 있다. 기술적 오류 관리에서도 두 플랫폼은 문제가 발생했을 때 사용자에게 이를 공지하고 해결 방법을 안내하며, 알림 기능을 통해 신속히 정보를 전달한다. 이러한 대응은 기술적 문제로 인해 사용자가 불편을 겪는 상황을 최소화하고자 하는 노력의 일환이다. 종합적으로 페이스북과 인스타그램은 기술적 오류 대응과 분쟁 해결에서 강점을 보이며, 사회적 책무를 이행하기 위한 체계적인 접근 방식을 보여주고 있다.

2) 실행 원칙

(1) 이용자를 위한 정보 공개

페이스북과 인스타그램은 정보 공개 측면에서 사용자에게 추천 및 배열 요소를 명확히 전달하는 선도적인 시스템을 운영하고 있다. 특히, 사용자는 앱 내에서 콘텐츠 추천 이유와 관련된 자신의 활동 데이터를 확인할 수 있으며, 이를 기반으로 더 많은 통제 옵션을 제공하여 사용자가 원하는 콘텐츠를 더 많이, 원하지 않는 콘텐츠를 더 적게 볼 수 있는 환경을 조성하고 있다.³⁰⁾ 이러한 접근은 사용자가 플랫폼의 AI 시스템을 더욱 신뢰할 수 있도록 정보를 명확히 제공하고, 사용자 경험을 개인화하며, 투명성을 강화하려는 페이스북과 인스타그램의 지속적인 노력을 보여준다고 할 수 있다.

(2) 이용자의 선택권 보장

페이스북과 인스타그램은 사용자가 자신의 선호와 필요에 따라 콘텐츠와 광고 타겟팅 설정을 직접 변경할 수 있는 기능을 제공하며, 선택권 보장에서 높은 수준의 접근성을 보장하고 있다. 예를 들어, 인스타그램 앱 내 설정에서는 사용자들이 광고 활동을 확인하고,

30) 메타, <https://about.fb.com/ko/news/2023/06/ai가-facebook과-instagram에서-보이는-콘텐츠에-영향을-미치는-방식/>

개별 광고를 숨기거나 신고할 수 있는 기능을 제공한다. 또한, 더 세부적인 광고 선호도 조정을 위해서는 페이스북 설정으로 이동하여 관심사, 고용주, 직책, 관계 상태 등을 기반으로 한 광고를 제한할 수 있다. 페이스북의 광고 선호 페이지에서 “광고 설정”을 통해 다른 웹사이트와 서비스에서 수집된 정보를 바탕으로 한 광고 표시를 중지할 수도 있다. 이러한 기능은 사용자가 플랫폼에서 경험하는 콘텐츠 배열과 광고 노출에 직접적인 영향을 미칠 수 있는 도구를 제공하며, 사용자 중심의 선택권을 보장하는 데 기여하고 있다.

(3) 자율검증 실행

페이스북과 인스타그램은 투명성 센터를 통해 글로벌 위험을 평가하고 이에 대비하기 위한 체계적인 자율검증을 실행하고 있다.³¹⁾ 특별히 위험 관리 과정에서 공개 보고서, 신뢰할 수 있는 파트너의 권장 사항, 콘텐츠 트렌드 관찰, 인권 실사, 정보팀의 평가 등을 통해 잠재적으로 유해한 트렌드를 식별하고, 이를 기반으로 알고리즘과 시스템을 조정한다. 또한, 위기 정책 프로토콜(CPP)은 콘텐츠 위험에 대응하고, 위험 증가 상황에서 알고리즘과 정책을 신속히 조정할 수 있는 체계를 제공한다. 이러한 조치는 페이스북과 인스타그램이 알고리즘 설계와 운영에서 투명성을 중시하고, 글로벌 위험 대응할 수 있는 자율검증 시스템을 갖추고 있음을 보여준다.

(4) 불만 처리 및 분쟁 해결

페이스북과 인스타그램은 불만 처리와 분쟁 해결에서 높은 수준의 신속성과 접근성을 제공하는 체계를 운영하고 있다. 신고 기능과 도움말 센터를 통해 사용자가 부적절한 콘텐츠나 계정을 쉽게 신고할 수 있으며, 신고된 문제는 플랫폼의 정책에 따라 신속히 처리된다. 무엇보다 사용자는 플랫폼 내에서 신고 기능과 도움말 센터에 쉽게 접근할 수 있어 불만 처리의 효율성을 높이고, 사용자 경험을 개선하는 데 기여하고 있다.

(5) 내부 규칙 제정

메타(페이스북과 인스타그램의 모회사)는 유럽연합(EU)의 DSA(Digital Services Act) 투명성 보고서를 통해 내부 규칙 제정과 콘텐츠 관리 방식에서 투명성을 강화하려는 노력을 보여주고 있다.³²⁾ 메타는 전 세계적으로 적용되는 커뮤니티 표준, 광고 표준, 상거래 정책

31) 메타, <https://transparency.meta.com/>

을 기반으로 자동화된 도구와 인공지능(AI)을 활용하여 정책 위반 콘텐츠와 계정을 감지하고 제거하는 체계를 운영하고 있다. AI는 대규모 콘텐츠를 실시간으로 분석하며, 인력 검토팀과 결합해 복잡한 맥락을 파악하고 공정성을 유지한다. 또한, 투명성 보고서를 통해 이러한 규칙과 관리 방식을 외부에 공개하며, 정책 실행의 책임성과 신뢰성을 제고하고 있다. 이러한 접근은 내부 규칙에 기반한 콘텐츠 관리 체계를 강화하는 데 기여하며, 메타가 사용자 신뢰를 확보하려는 의지를 명확히 보여준다. 다만, 내부 규칙과 알고리즘 설계의 구체적인 세부 내용을 더욱 널리 공개하고 외부 검증 가능성을 확대함으로써 투명성과 책임성을 더욱 강화할 필요가 있다.

바. 종합 분석 및 우수 사례

본 절에서는 넷플릭스, 디즈니+, 웨이브, 티빙과 같은 OTT 플랫폼과 페이스북 및 인스타그램과 같은 소셜 미디어 업체들이 추천 알고리즘의 핵심 원칙(투명성, 공정성, 책무성)과 실행 원칙(정보 공개, 선택권 보장, 자율검증 실행, 불만 처리 및 분쟁 해결, 내부 규칙 제정)을 얼마나 잘 이행하고 있는지를 이용약관과 규정을 통해 검증하고, 우수 사례 발굴 및 개선 방향 도출을 시도하였다. 분석 결과, 각 플랫폼은 원칙 구현 방식과 수준에서 차이가 드러났는데, 넷플릭스는 추천 이유를 사용자가 쉽게 이해하도록 메시지를 제공해 투명성을 높였으나, 디즈니+, 웨이브, 티빙은 데이터 활용 원칙을 설명하지만 실시간 알림이나 상세 설명은 부족한 상태였다. 공정성 측면에서 디즈니+는 연령 제한 및 가족 친화적 콘텐츠로 다양한 사용자 권익을 보호하며, 웨이브와 티빙은 딥러닝 기술을 통해 콘텐츠 편향성을 방지하고 있다. 넷플릭스는 글로벌 콘텐츠 큐레이션으로 공정성을 강화하고 있었다. 책무성 면에서는 넷플릭스가 기술적 오류 시정과 불만 처리 체계를 갖추고 있으며, 디즈니+는 부모 통제 기능으로 부적절한 콘텐츠 접근을 방지하고, 웨이브와 티빙은 데이터 보호 정책을 통해 책임 있는 플랫폼 운영을 강조하고 있음을 확인할 수 있었다.

한편 실행 원칙과 관련하여, 넷플릭스는 추천 알고리즘의 사용 사실과 이유를 사용자 친화적으로 명확히 알리고 있는 반면, 디즈니+와 웨이브는 데이터 활용 원칙은 공개하고 있으나 실시간 알림과 구체적인 설명 방식에서 미흡한 점이 있다. 디즈니+는 연령 제한과 콘텐츠 접근 권한 설정을 통해 선택권 보장에서 강점을 보이며, 웨이브와 티빙은 프로필

32) 메타, <https://transparency.meta.com/sr/dsa-transparency-report-oct2023-facebook/>

설정을 통해 개인화 콘텐츠를 관리할 수 있지만, 콘텐츠 배열 방식 변경 기능은 부족한 상황이다. 자율검증 실행 측면에서 웨이브는 테스트를 통해 알고리즘 성능을 평가하고 사용자 피드백을 반영해 시스템을 개선하고 있으며, 티빙도 알고리즘 효과를 정량적으로 분석하고 있으나 외부 검증 가능성은 부족하다. 디즈니+는 24시간 전화 상담과 실시간 채팅을 통해 사용자 불만을 신속히 해결하며 우수한 사례를 보여주고 있으며, 티빙은 다양한 지원 채널을 운영하나 실시간 알림 기능이 부족하다. 내부 규칙 제정에서는 넷플릭스와 디즈니+가 개인정보 보호와 알고리즘 관리 규정을 공개하며 투명성을 유지하고 있지만, 웨이브와 티빙은 내부 규정을 운영하되 외부 검증과 규정 공개가 부족해 개선이 필요해 보였다.

결론적으로, 넷플릭스, 디즈니+, 웨이브, 티빙 및 소셜 미디어 플랫폼들은 추천 알고리즘의 기본원칙과 실행 원칙을 다양한 수준에서 구현하고 있으며, 각 플랫폼 간의 차이가 뚜렷하게 나타났다. 넷플릭스는 투명성과 공정성 측면에서 우수한 사례를 보여주었고, 디즈니+는 가족 친화적 접근과 불만 처리에서 강점을 드러냈다. 웨이브와 티빙은 데이터 기반 추천과 자율검증에서 긍정적인 성과를 보였으나, 외부 검증과 규정 공개 측면에서 개선이 필요한 것으로 평가되었다. 이러한 분석은 각 플랫폼이 신뢰성과 사용자 경험을 더욱 강화하기 위해 어떤 부분을 보완해야 할지 방향성을 제시할 수 있을 것이다.

3. 설문조사

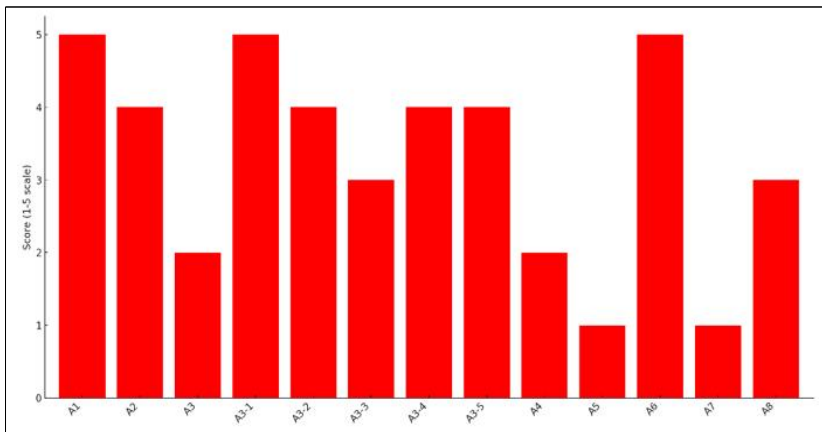
본 연구는 알고리즘 추천서비스 원칙의 이행 여부를 파악하기 위해 미디어 플랫폼 사업자를 대상으로 설문조사를 실시하였다. 총 2개 업체가 설문에 응답하였으며, 본 절에서는 이에 대한 정량적 결과를 제시하고자 한다. 이를 통해 기본원칙에 명시된 3가지 핵심 원칙(투명성, 공정성, 책무성)과 4가지 실행 원칙(정보 공개, 선택권 보장, 자율검증 실행, 불만 처리 및 분쟁 해결, 내부 규칙 제정)의 5개 항목 중 정보 공개 항목은 설문조사에서 제외됨)의 이행 실태를 파악하는 것이 목적이다. 응답 업체의 익명성 보장을 위해 두 업체를 각각 (A)기업과 (B)기업으로 표기한다.

가. 핵심 원칙

1) 투명성

(A)기업에 대한 알고리즘 활용 추천 서비스의 투명성과 관련된 설문조사는 A1(알고리즘 추천 서비스 제공 사실 알림), A2(알고리즘 추천 서비스 제공 사실 공지), A3(구체적인 추천 서비스 종류 공개), A3-1(관심 콘텐츠 제공 지원), A3-2(맞춤형 광고 제공), A3-3(맞춤형 이벤트 안내), A3-4(서비스 유지 및 개선에 데이터 활용), A3-5(서비스 변경 사항 알림), A4(실시간 문자나 알림을 통한 공지), A5(서비스 변경 사항과 이전 내용 공개), A6(이용자 개인정보 종류 및 범위 공지), A7(이해하기 쉬운 방식으로 개인정보 종류 및 범위 공지), A8(추천 서비스 종류별 개인정보 활용 유형 공개), 총 13개의 문항을 통해 측정되었다.

[그림 4-4] (A)기업 투명성 문항별 세부 점수



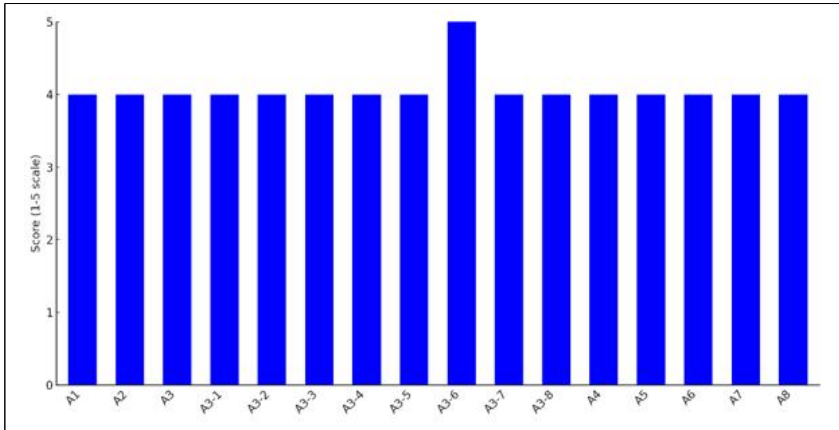
(A)기업은 A1(알고리즘 추천 서비스 제공 사실 알림)과 A2(알고리즘 추천 서비스 제공 사실 공지) 문항에 대해 각각 5점('매우 그렇다')과 4점('그런 편이다')으로 응답해 알고리즘 사용에 대한 기본적인 정보를 명확히 제공하고 있음을 나타냈다. A6(이용자 개인정보 종류 및 범위 공지) 항목에서도 마찬가지로 5점('매우 그렇다')으로 응답하여 (A)기업은 알고리즘 기반 추천 서비스의 제공 사실, 개인정보의 종류와 범위에 대한 정보를 제공하기 위해 노력하고 있는 것으로 나타났다.

하지만 구체적인 서비스 종류나 변경 정보 공개, 정보 제공 방식에서는 개선이 필요한 점이 드러났다. A3(구체적인 추천 서비스 종류 공개)의 세부 평가를 살펴보았을 때, 현재 제공되는 알고리즘 서비스의 구체적인 기능과 구현과 관련된 A3-1(관심 콘텐츠 제공 지원) 항목은 5점('매우 그렇다'), A3-2(맞춤형 광고 제공), A3-4(서비스 유지 및 개선에 데이터 활용), A3-5(서비스 변경 사항 알림) 항목은 각각 4점('그런 편이다')으로 높은 점수의 응답을 보여 실제 제공 기능과 효용성 측면에서 비교적 잘 운영되고 있는 것으로 나타났다. 그러나 A3(구체적인 추천 서비스 종류 공개) 항목에서 2점('그렇지 않은 편이다')이라는 상대적으로 낮은 점수의 응답을 보였고 A3-3(맞춤형 이벤트 안내)와 A4(실시간 문자나 알림을 통한 공지) 항목에서도 각각 3점('보통이다')과 2점('그렇지 않은 편이다')으로 낮은 점수를 보이며 정보 제공의 명확성과 접근성에 개선이 필요함을 시사하였다.

특히 A5(서비스 변경 사항과 이전 내용 공개)와 A7(이해하기 쉬운 방식으로 개인정보 종류 및 범위 공지)에 대한 응답은 각각 1점('전혀 그렇지 않다')으로 가장 낮은 점수의 응답을 기록하며 서비스 변경 사항에 대한 정보와 개인정보 공지 방식에서의 부족함을 드러냈다. 또한, A8(추천 서비스 종류별 개인정보 활용 유형 공개) 항목은 3점('보통이다')으로 나타나 개인정보 활용에 대한 구체적인 정보 공개는 다소 이루어지고 있지만, 여전히 개선이 필요한 수준임을 보여주었다. 이에 따라 (A)기업은 이용자와의 투명한 소통을 강화하고 개인정보 및 서비스 변경 사항에 대한 구체적이고 명확한 정보 제공을 위한 노력이 필요한 것으로 나타났다.

(B)기업에 대한 투명성 측정 문항은 A1(알고리즘 추천 서비스 제공 사실 알림), A2(알고리즘 추천 서비스 제공 사실 공지), A3(구체적인 추천 서비스 종류 공개), A3-1(이용자의 관심에 따라 흥미로운 콘텐츠 추천), A3-2(관심 있는 커뮤니티나 계정 추천), A3-3(관심 있는 다른 사용자 추천), A3-4(선호할 만한 제품 정보 제공), A3-5(맞춤형 광고 제공), A3-6(안전한 서비스 환경 제공), A3-7(기술적 문제 해결과 개선 노력), A3-8(서비스 변경 사항 알림), A4(실시간 문자나 알림을 통한 공지), A5(서비스 변경 사항과 이전 내용 공개), A6(이용자 개인정보 종류 및 범위 공지), A7(이해하기 쉬운 방식으로 개인정보 종류 및 범위 공지), A8(추천 서비스 종류별 개인정보 활용 유형 공개), 총 16개 문항으로 구성되었다.

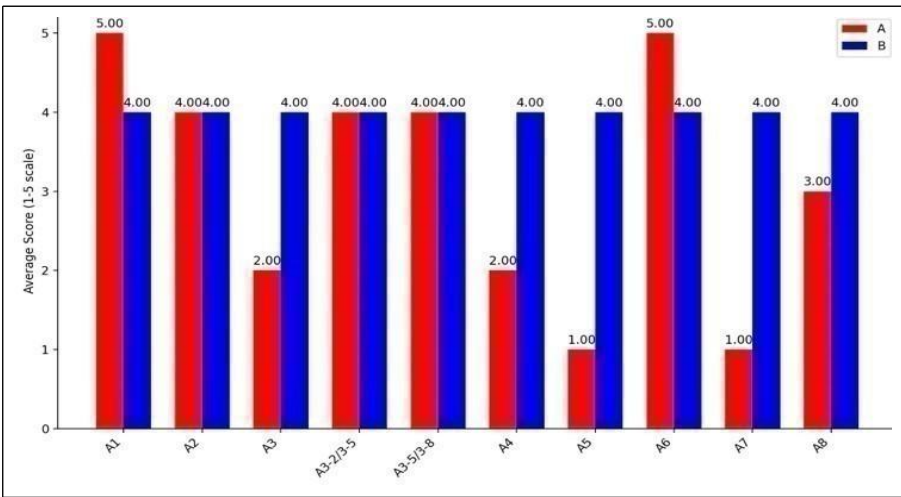
[그림 4-5] (B)기업 투명성 문항별 세부 점수



투명성 항목에 대한 분석 결과, (B)기업은 알고리즘 기반 추천 서비스의 제공 사실과 구체적인 서비스 종류, 개인정보 활용 범위, 변경 사항에 대한 공지 측면에서 체계적으로 운영되고 있다. A1(알고리즘 추천 서비스 제공 사실 알림)과 A2(알고리즘 추천 서비스 제공 사실 공지)에 대해 각각 4점('그런 편이다')으로 응답하여 알고리즘 사용에 대한 기본적인 정보를 명확히 제공하고 있음을 보여주었다. A4(실시간 문자나 알림을 통한 공지)와 A5(서비스 변경 사항과 이전 내용 공개) 역시 4점('그런 편이다')으로 응답되어, 이용자와의 소통 방식과 변경 정보 제공에서 긍정적인 노력을 기울이고 있음을 나타냈다. A3(구체적인 추천 서비스 종류 공개) 관련 항목에서는 A3-1(이용자의 관심에 따라 흥미로운 콘텐츠 추천), A3-2(관심 있는 커뮤니티나 계정 추천), A3-3(관심 있는 다른 사용자 추천), A3-4(선호할 만한 제품 정보 제공), A3-5(맞춤형 광고 제공), A3-7(기술적 문제 해결과 개선 노력), A3-8(서비스 변경 사항 알림) 모두 4점('그런 편이다')으로 응답되어 추천 서비스의 구체적인 기능을 비교적 높은 수준으로 제공하고 있음을 보여주었다. 특히 A3-6(안전한 서비스 환경 제공)은 5점('매우 그렇다')으로 응답되어 이용자 안전과 관련된 서비스 개선을 위해 적극적으로 노력하고 있는 것으로 나타났다. 또한, A6(이용자 개인정보 종류 및 범위 공지), A7(이해하기 쉬운 방식으로 개인정보 종류 및 범위 공지), A8(추천 서비스 종류별 개인정보 활용 유형 공개) 항목에서도 모두 4점('그런 편이다')으로 응답되어, 개인정보 활용과 관련된 정보 제공에서 꾸준한 노력을 기울이고 있음을 나타냈다.

설문 결과는 (B)기업이 알고리즘 기반 추천 서비스 운영에서 투명성을 중시하며 이용자와의 소통과 명확한 정보 제공 측면에서 긍정적으로 평가하고 있음을 보여준다. 따라서 전반적으로 (B)기업은 이용자들에게 구체적인 정보와 변경 사항을 명확히 전달하고 있으며 특히 이용자 안전과 관련된 서비스 개선을 위해 노력하고 있는 것으로 나타났다. 이는 (B)기업이 투명성을 유지하기 위해 공지와 설명 방식을 체계적으로 마련한 사례로 볼 수 있다.

[그림 4-6] (A)기업과 (B)기업의 투명성 문항별 세부 점수 비교



위 그래프는 (A)기업과 (B)기업의 투명성 세부 항목(A1~A8 및 A3의 공유된 질문 항목)을 비교한 결과를 보여준다. A1(알고리즘 추천 서비스 제공 사실 알림)에서는 (A)기업이 5점, (B)기업이 4점으로 응답하며 (A)기업이 해당 항목에서 강점을 보였다. A2(알고리즘 추천 서비스 제공 사실 공지)는 두 기업 모두 4점으로 동일한 응답을 보여 기본적인 정보 제공에 있어 유사한 노력을 기울이고 있음을 나타냈다. A3(구체적인 추천 서비스 종류 공개)의 경우, (A)기업이 2점, (B)기업이 4점으로 응답하여 (B)기업이 구체적인 추천 서비스의 종류를 공개하는 데 있어 더 높은 수준의 정보를 제공하고 있음을 보여준다. A3-2 및 A3-5 (맞춤형 광고 제공)와 A3-8 (서비스 변경 사항 알림)에서는 두 기업 모두 4

점으로 동일한 응답을 보이며 관련 정보 제공에서 비슷한 수준의 노력을 기울이고 있음을 확인하였다.

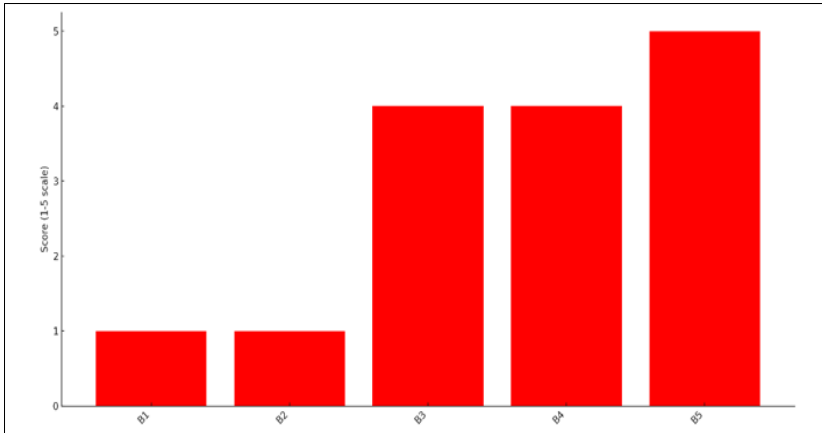
A4(실시간 문자나 알림을 통한 공지)에서는 (A)기업이 2점, (B)기업이 4점으로 응답하여 (B)기업이 이용자와의 실시간 소통 체계에서 더 강점을 보였다. A5(서비스 변경 사항과 이전 내용 공개) 항목에서는 (A)기업이 1점, (B)기업이 4점으로 응답함으로써 (B)기업이 이 영역에서 훨씬 더 높은 수준의 투명성을 보이고 있음을 나타냈다. A6(이용자 개인정보 종류 및 범위 공지)에서는 (A)기업이 5점으로 응답하며 (B)기업의 4점보다 높은 점수를 기록하여 개인정보의 종류와 범위 공지에서 강점을 드러냈다. 그러나 A7(이해하기 쉬운 방식으로 개인정보 종류 및 범위 공지)에서는 (A)기업이 1점으로 낮은 응답을 보인 반면, (B)기업은 4점으로 응답하여 이용자 친화적인 정보 전달 방식에서 (B)기업이 더 우위를 보였다. A8(추천 서비스 종류별 개인정보 활용 유형 공개)에서도 (A)기업이 3점, (B)기업이 4점으로 응답하며 (B)기업이 개인정보 활용 유형 공개에서 상대적으로 더 나은 결과를 보였다.

전반적으로, (B)기업은 투명성 영역의 대부분의 항목에서 (A)기업보다 더 높은 응답을 보이며 구체적이고 명확한 정보 제공에 강점을 보였다. 반면, (A)기업은 일부 항목에서 강점을 보였으나 특정 세부 항목에서는 이용자 중심의 정보 제공 방식과 서비스 변경 사항 공개와 관련하여 개선이 필요한 것으로 나타났다.

2) 공정성

B1(알고리즘 결과 기준 공개), B2(추천 결과의 근거 공지), B3(알고리즘의 영향 사전 평가), B4(알고리즘의 영향 사후 평가), B5(시스템 지속적 개선)은 (A)기업에 대한 공정성 측정을 위해 사용된 설문지 문항이다.

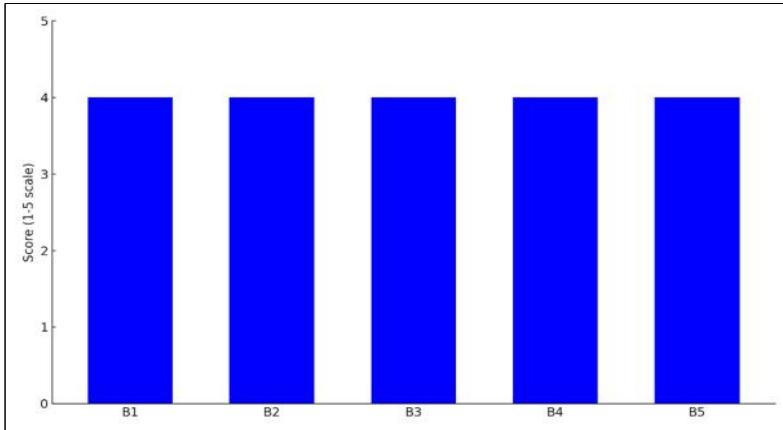
[그림 4-7] (A)기업 공정성 문항별 세부 점수



분석 결과, (A)기업은 공정성 확보를 위해 사전, 사후 평가 및 지속적인 시스템 개선 노력을 기울이고 있으나 알고리즘 결과의 기준과 추천 근거를 이용자에게 공개하지 않는다는 점에서 정보 공개의 측면에서 한계를 보인다. B3(알고리즘의 영향 사전 평가), B4(알고리즘의 영향 사후 평가) 항목은 각각 4점('그런 편이다'), B5(시스템 지속적 개선) 항목에서는 5점('매우 그렇다')의 응답을 보여 서비스의 평가 및 개선을 위한 지속적인 노력이 잘 이루어지고 있음을 나타냈다. 하지만 B1(알고리즘 결과 기준 공개)과 B2(추천 결과의 근거 공지) 항목은 각각 1점('전혀 그렇지 않다')으로 매우 낮은 응답을 기록하며 알고리즘 결과와 근거를 이용자에게 명확히 공개하는 데 상당한 개선이 필요한 것으로 드러났다. 이는 공정성을 강화하기 위해 알고리즘의 작동 방식과 기준을 투명하게 설명하는 구체적인 정보 공개가 필요하다는 점을 시사한다.

(B)기업에 대한 공정성 항목 측정 문항 역시 동일하게 B1(알고리즘 결과 기준 공개), B2(추천 결과의 근거 공지), B3(알고리즘의 영향 사전 평가), B4(알고리즘의 영향 사후 평가), B5(시스템 지속적 개선), 5개 문항으로 구성되었다.

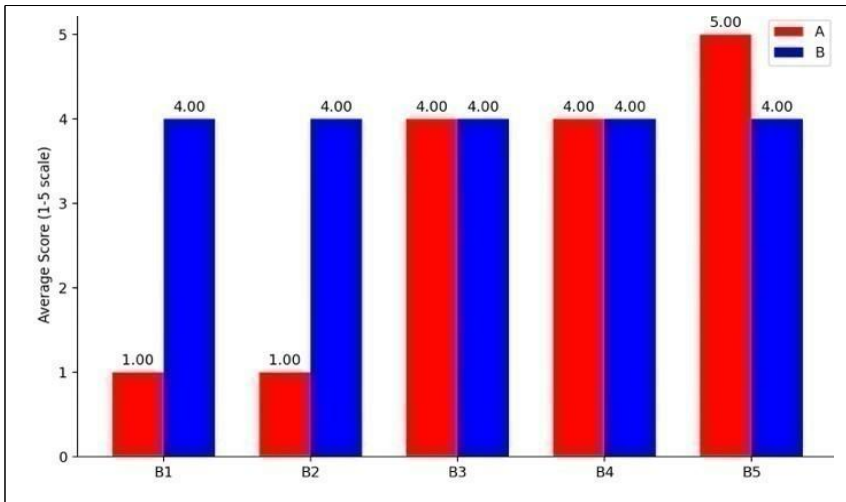
[그림 4-8] (B)기업 공정성 문항별 세부 점수



설문 결과에 따르면 B1(알고리즘 결과 기준 공개), B2(추천 결과의 근거 공지), B3(알고리즘의 영향 사전 평가), B4(알고리즘의 영향 사후 평가), B5(시스템 지속적 개선) 문항 모두 4점(그런 편이다)으로 응답되며 (B)기업이 이용자에게 공정성을 보장하기 위해 지속적인 노력을 기울이고 있음을 보여준다. 알고리즘 결과의 기준과 근거를 명확히 공지하고, 추천 서비스의 영향을 사전적으로 평가하며, 사후적으로도 지속적으로 관리하고 있음을 확인하였다. 또한, 시스템 개선을 통해 공정성을 강화하는 데도 꾸준히 힘쓰고 있는 것으로 나타났다.

전반적으로, (B)기업은 알고리즘 추천 서비스의 공정성을 확보하기 위해 적절한 노력을 기울이고 있으며, 이용자 정보 공개, 사전·사후적 영향 평가, 시스템 개선 등 모든 측면에서 체계적인 접근 방식을 채택하고 있다. 알고리즘 추천 서비스의 공정성을 유지하기 위해 지속적인 조치를 시행하고 있으며 이는 기업이 공정성을 유지하기 위한 기본적인 노력이 이루어지고 있음을 보여준다.

[그림 4-9] (A)기업과 (B)기업의 공정성 문항별 세부 점수 비교



(A)기업과 (B)기업의 공정성 세부 항목(B1~B5)에 대한 비교 결과, B1(알고리즘 결과 기준 공개)에서는 (A)기업이 1점, (B)기업이 4점으로 응답하며 (B)기업이 알고리즘 결과 기준에 대한 정보 공개에서 더 높은 수준의 투명성을 보여주었다. B2(추천 결과의 근거 공지)에서도 (A)기업이 1점, (B)기업이 4점으로 응답하여 추천 결과의 근거를 명확히 전달하는 데 있어 (B)기업이 더 체계적인 노력을 기울이고 있음을 나타냈다.

B3(알고리즘의 영향 사전 평가)와 B4(알고리즘의 영향 사후 평가)에서는 두 기업 모두 4점으로 응답하며 알고리즘이 미치는 영향을 사전에 평가하고 이후에도 지속적으로 검토하는 데 있어 비슷한 수준의 노력을 기울이고 있음을 보여주었다. B5(시스템 지속적 개선)에서는 (A)기업이 5점, (B)기업이 4점으로 응답하며 (A)기업의 시스템 개선을 통해 공정성을 강화하려는 강한 의지를 확인하였다. 이는 (A)기업이 서비스 개선과 공정성 확보를 기업 운영의 중요한 요소로 인식하고 있음을 시사하는 결과이다.

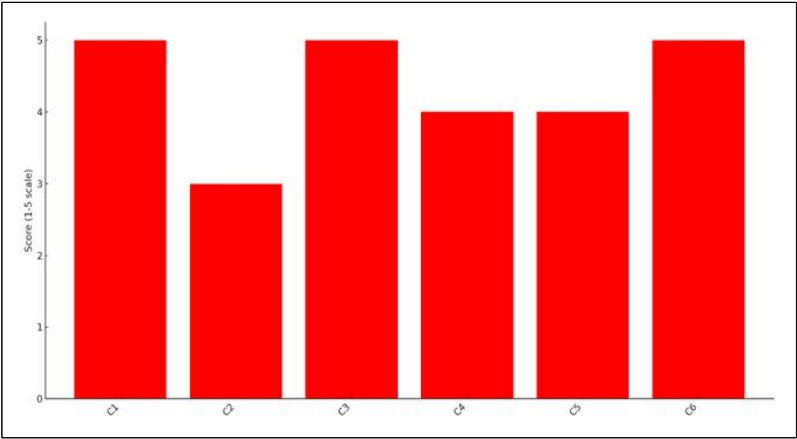
전반적으로, (B)기업은 알고리즘 결과 기준 공개와 추천 결과의 근거 공지에서 높은 점수의 응답을 보이며 정보 공개와 투명성에서 강점을 보였다. 반면, (A)기업은 시스템 지속적 개선에서 더 높은 응답을 보이며 시스템 개선을 통한 공정성 확보에 주력하고 있음을 나타냈다. 두 기업은 알고리즘의 사전, 사후 영향 평가에서는 비슷한 수준의 노력을 보이

고 있으며 이외 공정성 항목에서 각기 다른 강점을 가지고 있는 것으로 분석된다.

3) 책무성

C1(투명성과 공정성 향상 노력), C2(투명성 및 공정성 향상 노력 사항 공지), C3(기능적 오류나 오작동 발생 시 시정 노력), C4(문제 상황 및 시정 신속한 공지), C5(문제 상황 및 시정 누락 없는 공지), C6(이용자 불만 해결 노력) 문항은 (A)기업의 책무성 설문 문항으로 사용되었다.

[그림 4-10] (A)기업 책무성 문항별 세부 점수



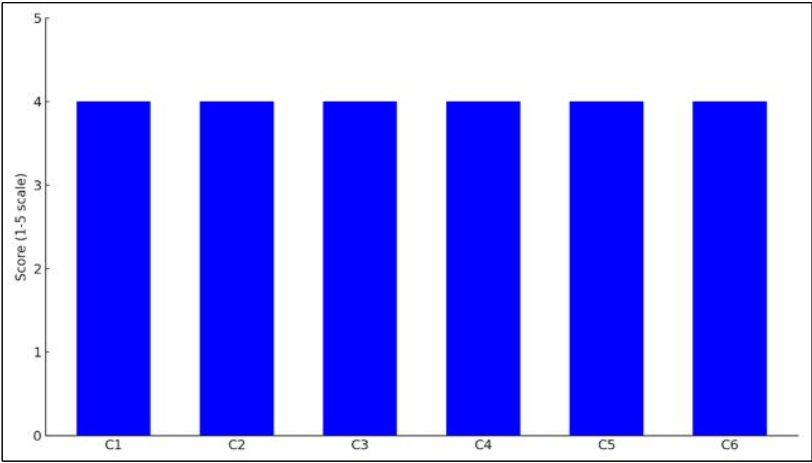
(A)기업은 알고리즘 추천 서비스의 투명성과 공정성 강화를 위해 노력하며, 오류 및 문제 발생 시 이를 신속히 시정하고 이용자에게 공지하는 체계를 구축하고 있는 것으로 나타났다. C1(투명성과 공정성 향상 노력) 문항은 5점(‘매우 그렇다’)으로 응답되었고 투명성과 공정성을 기술적으로 강화하기 위해 추가적으로 “Explainable AI 기술 개발 예정”이라는 계획을 언급하며 구체적인 노력을 보였다. C3(기능적 오류나 오작동 발생 시 시정 노력) 문항 역시 5점(‘매우 그렇다’)으로 응답되었으며 문제 상황에 체계적으로 대처하기 위해 “장애 대응팀 운영”이 이루어지고 있음을 강조하였다. C4(문제 상황 및 시정 신속한 공지)와 C5(문제 상황 및 시정 누락 없는 공지) 문항은 각각 4점(‘그런 편이다’)으로 문제 해결 상황을 빠르고 정확하게 전달하려는 노력이 이루어지고 있음을 보여주었다. C6(이용자 불

만 해결 노력)은 5점(‘매우 그렇다’)으로 응답되었으며, “고객센터 FAQ 및 1:1 고객문의 운영, 이슈 접수 시 빠른 개발 대응으로 이슈 해결”을 통해 이용자와의 신속한 소통과 문제 해결 노력이 드러났다.

다만, 이러한 문제 상황에 대한 시정 및 공지 노력은 이루어지고 있으나 투명성과 공정성 향상을 위한 노력을 이용자에게 명확히 전달하지 않는 부분에서 개선이 필요한 점이 나타났다. C2(투명성 및 공정성 향상 노력 사항 공지) 항목은 3점(‘보통이다’)으로 응답되어, “이용자를 위한 플랫폼 서비스 고도화를 위해 지속적 노력, 대외 커뮤니케이션 진행”과 같은 활동을 통해 소통을 강화하려는 노력을 기울이고 있음을 밝혔으나 공지 측면에서 다소 부족한 면이 드러났다. 전반적으로 해당 기업은 책무성 항목에서 평균 4.3점을 보이며 책임감을 바탕으로 문제 해결 및 공지 체계를 운영함과 동시에 투명성과 공정성을 강화하기 위한 노력을 기울이고 있음을 알 수 있었다.

(B)기업은 C1(투명성과 공정성 향상 노력), C2(투명성 및 공정성 향상 노력 사항 공지), C3(기능적 오류나 오작동 발생 시 시정 노력), C4(문제 상황 및 시정 신속한 공지), C5(문제 상황 및 시정 누락 없는 공지), C6(이용자 불만 해결 노력) 및 주관식 문항으로 구성된 책무성 항목에 대하여 모든 항목에 4점(‘그런 편이다’)으로 응답하며 투명성과 공정성 향상, 문제 해결 및 공지 체계를 책임감 있게 운영하고 있음을 보여주었다.

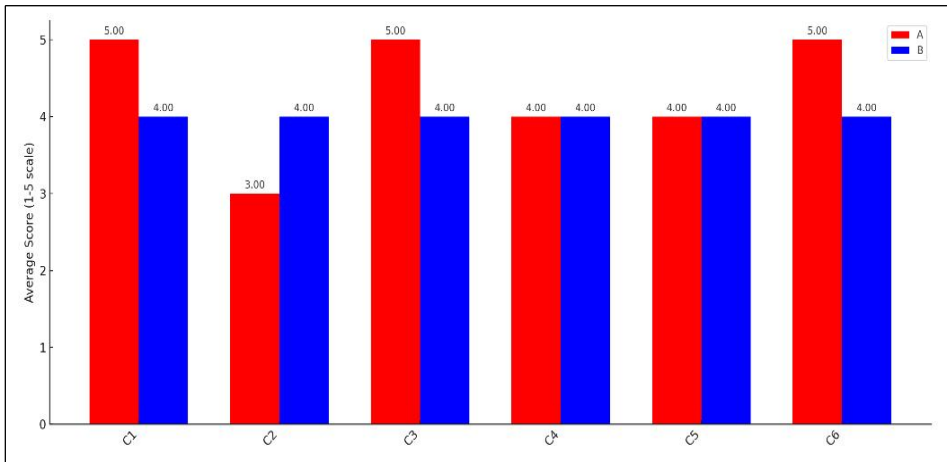
[그림 4-11] (B)기업 책무성 문항별 세부 점수



C1(투명성과 공정성 향상 노력)과 관련하여 “사용자 컨트롤 및 맞춤 설정을 제공하고, 공익 연구를 지원하기 위해 콘텐츠 라이브러리와 API를 제공합니다. 그리고 추천 가이드 라인 또한 발표했습니다”라는 응답을 통해 투명성과 공정성을 강화하기 위한 구체적인 노력을 나타냈다. C2(투명성 및 공정성 향상 노력 사항 공지) 항목에 대해 “자사의 Newsroom, AI System Cards, Transparency 등 다양한 웹사이트를 통해 알리고 있습니다”라는 답변을 통해 이용자와의 소통을 강화하려는 모습을 보였다.

C3(기능적 오류나 오작동 발생 시 시정 노력) 항목에 추가적으로 “기술적 오류 혹은 업데이트가 있을 시 자사의 Newsroom에 공지하고 있습니다”라고 응답하여 문제 상황 발생 시 신속히 공지하는 체계를 마련하고 있음을 확인할 수 있었다. C4(문제 상황 및 시정 신속한 공지)와 C5(문제 상황 및 시정 누락 없는 공지) 항목에서는 문제 해결 상황을 투명하게 전달하려고 노력하고 있음을 보여주었으며, C6(이용자 불만 해결 노력)에 대해서는 “문제가 발견되었을 때 이용자가 직접 신고할 수 있는 문제신고 기능을 각 플랫폼에서 제공하고 있습니다”라는 응답을 통해 이용자와의 소통과 문제 해결을 위한 체계적 방안을 마련하고 있음을 나타냈다. 전반적으로, 해당 기업은 모든 책무성 항목에서 균일하게 높은 수준의 노력을 기울이고 있으며 투명성과 공정성을 기반으로 이용자 신뢰를 향상시키기 위한 체계를 유지하고 있음을 보여준다.

[그림 4-12] (A)기업과 (B)기업의 책무성 문항별 세부 점수 비교



위 그래프는 (A)기업과 (B)기업의 책무성 세부 항목(C1-C6)에 대한 비교를 보여준다. C1(투명성과 공정성 향상 노력)에서는 두 플랫폼 모두 4점 이상으로 응답하여 투명성과 공정성을 강화하려는 노력이 확인되었으며, 특히 (A)기업은 5점으로 응답하여 투명성 및 공정성 향상을 위해 적극적으로 노력하고 있음이 확인되었다. 주관식 응답에서 (A)기업은 Explainable AI 기술 개발을 통해 기술적 발전을 이루고, 이를 기반으로 알고리즘 추천 서비스의 투명성과 공정성을 더욱 강화할 계획임을 밝혔다. 반면, (B)기업은 사용자에게 맞춤 설정과 컨트롤 기능을 제공하며, 공익 연구를 지원하기 위해 콘텐츠 라이브러리와 API를 활용하고, 추천 서비스의 투명성을 높이기 위한 가이드라인을 발표하는 등 이용자의 자율성과 공익을 고려한 접근을 강조하였다.

C2(투명성 및 공정성 향상 노력 사항 공지)에서는 (B)기업이 4점으로 (A)기업의 3점보다 약간 높은 점수의 응답을 보이며 공지 사항 전달력에 있어서 다소 앞선 모습을 보였다. 특히, (A)기업은 이용자를 위한 플랫폼 서비스의 고도화를 위해 지속적으로 노력하고 있으며 외부와의 소통을 강화하기 위한 대외 커뮤니케이션 활동을 진행하고 있다. 반면, (B)기업은 자사의 다양한 웹사이트를 통해 정보를 제공하며 이용자와의 소통을 이어가고 있는 것으로 나타났다. C3(기능적 오류나 오작동 발생 시 시정 노력)에서는 (A)기업이 5점으로 (B)기업의 4점보다 더 높은 점수의 응답을 보이며 기술적 문제 대응 체계에서 강점을 보였다. 보다 구체적으로는 (A)기업은 장애 대응팀을 운영하며 기술적 문제 발생 시 신속하게 대처할 수 있는 체계를 갖추고 있다. 반면, (B)기업은 기술적 오류나 업데이트가 발생할 경우 자사의 Newsroom을 통해 이를 공지하며 문제 상황을 해결하기 위한 시정 노력을 기울이고 있다.

C4(문제 상황 및 시정 신속한 공지)와 C5(문제 상황 및 시정 누락 없는 공지) 항목에서는 두 플랫폼 모두 4점으로 응답하며 이용자와의 투명한 소통을 위한 노력이 돋보였다. C6(이용자 불만 해결 노력)에서는 (A)기업이 5점으로 (B)기업의 4점보다 약간 높은 응답을 보이며 이용자 불만 처리와 문제 해결에서 강한 성과를 보였다. (A)기업은 고객센터를 통해 FAQ와 1:1 고객문의 서비스를 운영하며, 이슈 접수 시 빠른 개발 대응을 통해 문제를 해결하는 체계를 갖추고 있으며, (B)기업은 문제가 발생했을 때 이용자가 직접 신고할 수 있는 기능을 각 플랫폼에서 제공하며 이용자 주도의 문제 해결을 지원하고 있다.

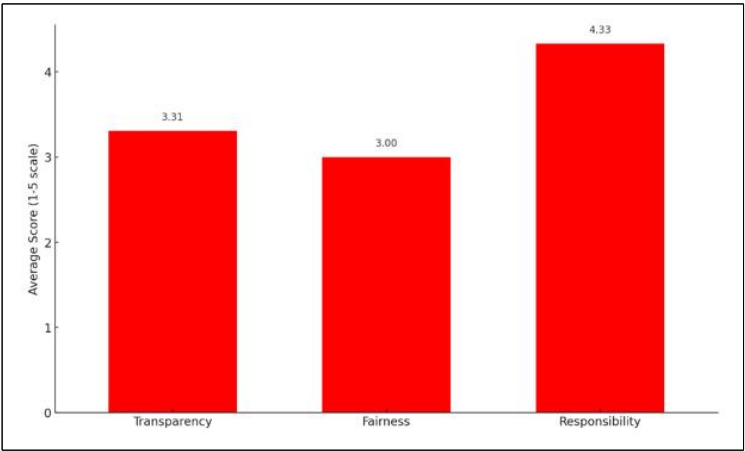
전반적으로 두 플랫폼 모두 책임감 있는 운영을 강조하고 있으나, (A)기업이 전반적으로

약간 더 높은 점수의 응답을 보이며 책무성 측면에서 비교적 강점을 보이고 있음을 알 수 있다.

4) 핵심 원칙 항목 비교 평가

핵심 원칙 항목인 투명성(Transparency), 공정성(Fairness), 책무성(Responsibility)에 대한 (A)기업의 평균 점수를 비교한 결과, 책무성 항목이 평균 4.33점으로 가장 높은 점수를 보여 책임감 있는 서비스 운영과 관련된 활동에서 강점을 보이고 있는 것으로 나타났다. 반면, 투명성 항목은 평균 3.31점으로 상대적으로 낮은 점수로 응답되며 이용자와의 정보 공유와 명확한 커뮤니케이션에서 개선이 필요한 영역임을 드러냈다. 공정성 항목은 평균 3 점으로, 공정한 알고리즘 작동과 이용자 신뢰 형성을 위한 추가적인 노력이 요구되는 것으로 나타났다. 이는 전반적으로 (A)기업은 책무성 측면에서 강점을 보이는 반면, 투명성과 공정성에서는 개선이 필요한 부분이 있음을 시사한다.

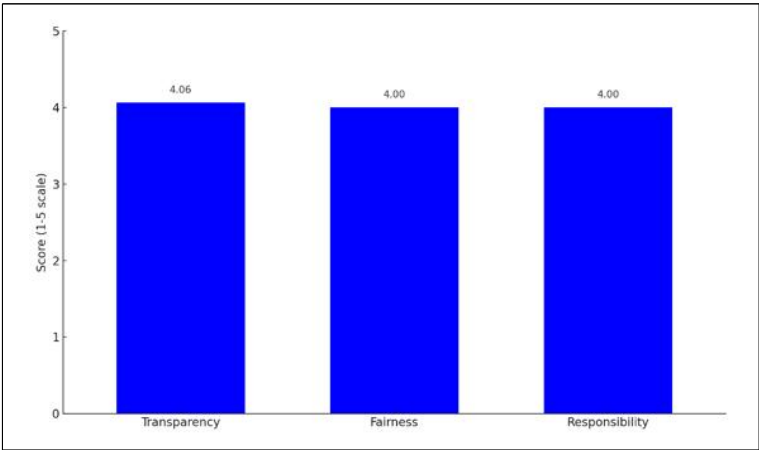
[그림 4-13] (A)기업 핵심 원칙 항목별 평균 비교



(B)기업의 투명성, 공정성, 책무성의 평균 점수를 비교한 결과를 보여준다. 투명성 (Transparency)은 평균 4.06점으로, (B)기업이 알고리즘 사용에 대한 정보 공개와 서비스의 구체적인 내용을 이용자에게 알리는 데 적절하게 노력하고 있음을 나타낸다. 공정성 (Fairness)은 평균 4점으로, 추천 결과의 기준과 근거 공개, 영향 평가와 관련된 부분에서

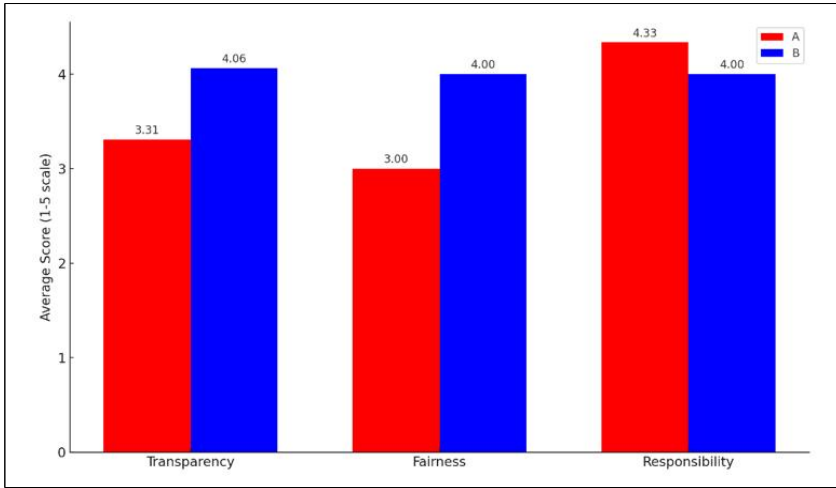
안정적인 성과를 보이고 있다. 책무성(Responsibility) 항목 또한 평균 4점으로, 문제 시정 및 공지 체계를 통해 이용자 신뢰를 강화하고 있으며, 책임감 있는 운영이 이루어지고 있음을 보여준다.

[그림 4-14] (B)기업 핵심 원칙 항목별 평균 비교



전반적으로, 해당 기업은 투명성, 공정성, 책무성의 핵심 원칙을 기반으로 알고리즘 추천 서비스를 신뢰성 있게 운영하고 있다. 특히 투명성 측면에서는 이용자에게 서비스 작동 사실과 이용되는 이용자 개인정보의 종류와 범위를 명확히 전달하며, 알고리즘 작동 방식에 대한 이용자의 이해를 높이는 데 기여하고 있다. 공정성 면에서는 추천 결과의 객관성과 신뢰성을 확보하기 위해 편향성을 줄이고 결과의 명확한 설명을 통해 안정적인 평가를 받고 있다. 책무성 항목에서도 문제 해결과 공지 체계 운영에 있어 강점을 보이고 있으며 이러한 노력은 이용자와의 소통을 강화하고 서비스 운영의 신뢰성을 높이는 데 기여하고 있다.

[그림 4-15] 핵심 원칙 항목별 (A)기업과 (B)기업의 평균 비교



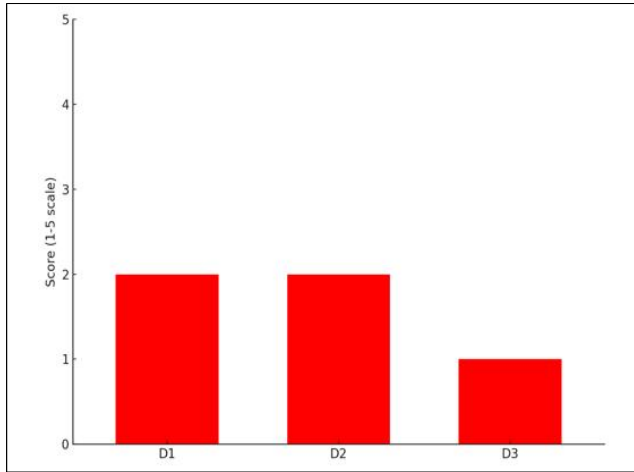
위 그래프는 (A)기업과 (B)기업의 주요 핵심 원칙인 투명성, 공정성, 책무성에 대한 설문 조사 결과를 비교한 것이다. 투명성 항목에서는 (B)기업이 평균 4.06점으로 (A)기업의 3.31점보다 높은 응답을 보여 알고리즘 관련 정보 공개와 이용자와의 소통 면에서 (B)기업이 더 나은 성과를 보이고 있음을 나타냈다. 공정성에서는 (A)기업과 (B)기업이 각각 평균 3점과 4점으로 응답되어, 추천 결과의 기준과 근거 공개에서 (B)기업이 상대적으로 더 적극적으로 노력하고 있음을 나타냈다. 책무성에서는 (A)기업과 (B)기업이 각각 4.33점과 4점으로 응답되어, 두 기업 모두 책임감 있는 운영을 보이고 있으나 (A)기업이 약간 더 높은 응답을 보였다. 전반적으로 (B)기업은 투명성과 공정성에서 더 긍정적으로 평가하였으며, (A)기업은 책무성 측면에서 강점을 보이고 있음을 확인할 수 있다.

나. 실행 원칙

1) 이용자의 선택권 보장

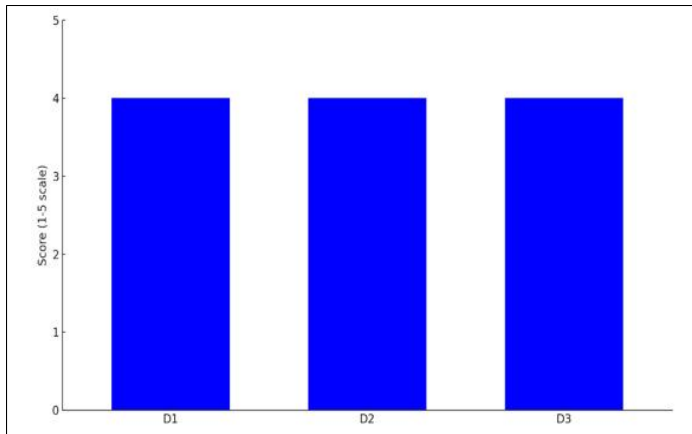
이용자 선택권 보장 항목에서는 D1(알고리즘 추천 서비스 변경 허용), D2(서비스 제공 정도 변경 방법 안내), D3(변경 시 서비스 이용 변화에 대한 안내) 세부 항목이 포함되었다.

[그림 4-16] (A)기업 이용자 선택권 문항별 세부 점수



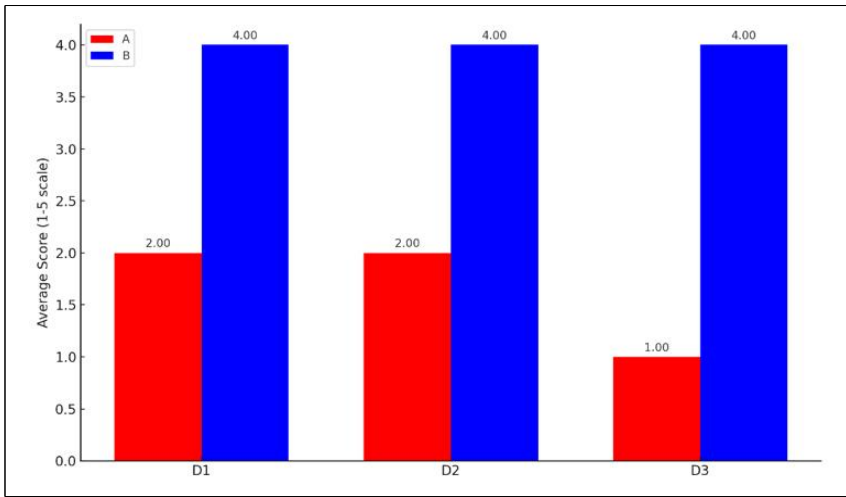
(A)기업의 이용자 선택권 보장 수준은 다른 기본원칙에 비해 상대적으로 부족한 것으로 나타났다. 이용자 보호 기본원칙 문항 가운데 평균 1.7점으로 가장 낮은 점수의 응답을 보였으며 모든 설문 문항에서 부정적인 응답이 나타났다. D1(알고리즘 추천 서비스 변경 허용)과 D2(서비스 제공 정도 변경 방법 안내) 문항은 각각 2점(‘그렇지 않은 편이다’)으로, 이용자에게 선택권 조정의 가능성과 방법에 대한 안내가 충분히 이루어지지 않음을 시사했다. D3(변경 시 서비스 이용 변화에 대한 안내) 문항은 1점(‘전혀 그렇지 않다’)으로 나타나, 서비스 변경 사항에 따른 영향을 전혀 전달하지 못하고 있음을 보여주었다. 전반적으로, (A)기업은 이용자 선택권 보장을 위한 조치가 부족한 것으로 보이며, 알고리즘 활용 추천 서비스를 제공받는 정도의 변경과 그 구체적인 방식, 그로 인한 영향을 이용자에게 명확히 안내할 필요가 있는 것으로 나타났다.

[그림 4-17] (B)기업 이용자 선택권 문항별 세부 점수



(B)기업에 대한 선택권 항목은 D1(알고리즘 추천 서비스 변경 허용), D2(서비스 제공 정도 변경 방법 안내), D3(변경 시 서비스 이용 변화에 대한 안내), 3개의 세부 문항을 통해 측정되었다. 이용자 선택권 모든 항목에서 (B)기업은 4점(‘그런 편이다’)으로 응답하여 이용자의 자율성을 보장하기 위해 적극적으로 노력하고 있음을 보여주었으며 이용자의 선택권 보장 측면에서 강점을 나타냈다. 전반적으로 (B)기업은 이용자가 자신의 필요와 선호에 따라 추천 서비스를 제공받는 정도를 변경할 수 있도록 설계하고, 변경 방법과 변경에 따른 영향을 이해하기 쉽게 안내하는 등 선택권 보장을 충실히 이행하고 있는 것으로 평가된다. 선택권 보장의 구체적 구현과 이용자 경험을 더욱 명확하게 안내함으로써 해당 기업은 이 영역에서 더 높은 성과를 달성할 수 있을 것으로 기대된다.

[그림 4-18] (A)기업과 (B)기업의 선택권 문항별 세부 점수 비교



(A)기업과 (B)기업의 이용자 선택권 보장 항목(D1~D3)에 대한 비교 결과, D1(알고리즘 추천 서비스 변경 허용)에서는 (A)기업이 2점, (B)기업이 4점으로 응답하며 (B)기업이 이용자에게 알고리즘 추천 서비스를 제공받는 정도를 조정할 수 있는 선택권을 보다 적극적으로 제공하고 있음을 나타냈다. D2(서비스 제공 정도 변경 방법 안내)에서도 마찬가지로 (A)기업은 2점, (B)기업은 4점을 기록하며 변경 방법에 대한 정보 제공에 있어서 (B)기업이 더 체계적인 모습을 보였다. D3(변경 시 서비스 이용 변화에 대한 안내) 항목에서는 (A)기업이 1점, (B)기업이 4점을 응답하며 변경에 따른 서비스 이용 변화에 대한 명확한 안내에서도 (B)기업이 우위를 보였다.

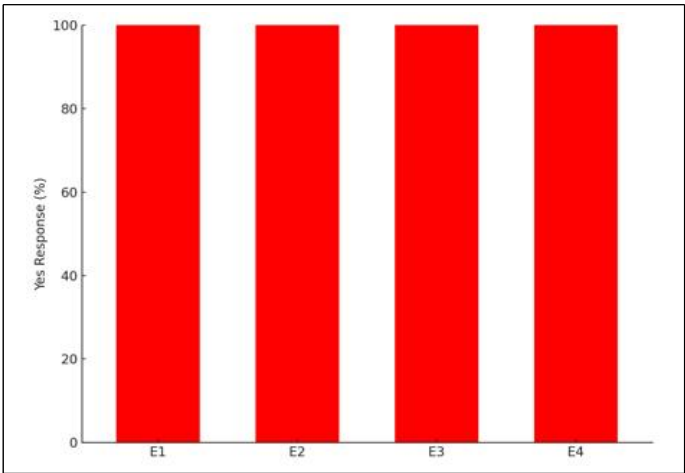
전반적으로, (B)기업은 이용자 선택권 보장 항목에서 체계적이고 명확한 정보를 제공하며 이용자 중심의 설계를 갖추고 있음을 보여주었다. 반면, (A)기업은 선택권 제공과 관련된 체계와 정보 제공 면에서 부족함을 보이며 개선이 요구되는 영역임을 시사하였다.

2) 자율검증 실행

E1(자율검증 체계 구축), E2(체계 확인 및 개선), E3(체계 기록 및 보관), E4(체계 공개)로 구성된 자율검증 실행 항목에 대한 (A)기업의 설문 결과를 분석한 결과, 모든 항목에서 100%의 ‘예’ 응답 비율을 보이며, 해당 기업이 자율검증 실행 원칙을 충실히 이행하고 있

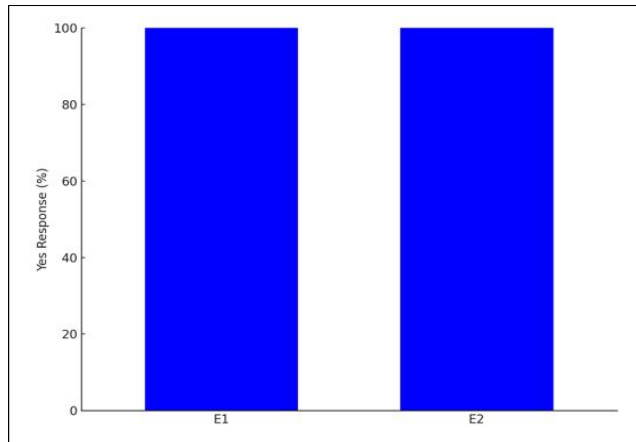
음을 나타냈다.

[그림 4-19] (A)기업 자율검증 문항별 세부 점수



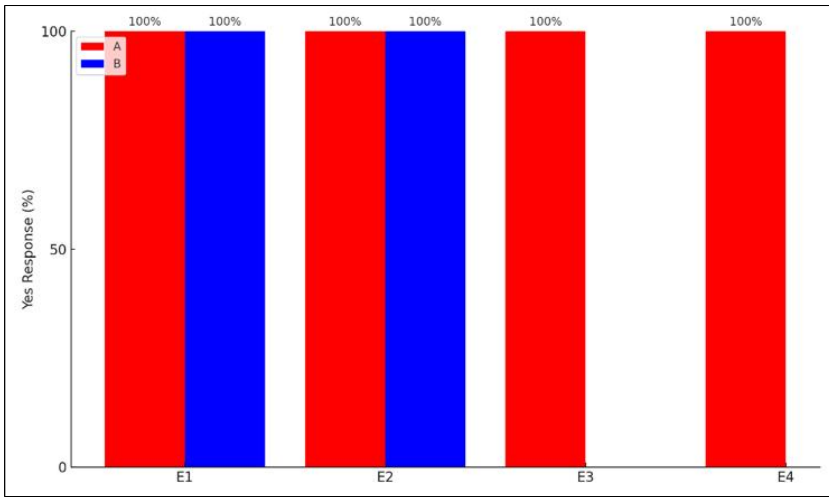
특히 (A)기업은 “추천 관련 정기 회의에서 검증”하는 방식으로 알고리즘 활용 추천 서비스가 이용자에게 미치는 잠재적인 위험을 예측하고 평가하며 상시적으로 위험을 관리하는 자율검증 체계를 운영하고 있을 뿐만 아니라, “회사 내부의 누구나 회의록 열람이 가능”하도록 하여 누구나 쉽게 자율검증 체계를 열람할 수 있도록 공개되고 있다. 이는 (A)기업이 알고리즘의 잠재적 위험을 지속적으로 관리하고 자율검증 체계를 확립 및 운영하는 데 있어 강한 책임감을 보여주는 것으로 평가된다. 따라서 (A)기업은 자율검증 체계를 전반적으로 체계적으로 운영하고 있다. 추천 서비스의 위험 관리를 위한 체계를 구축하고, 이를 통해 서비스 운영 과정에서 발생할 수 있는 문제를 관리 및 개선하며, 기록과 보관을 통해 자율검증에 활용할 수 있도록 하고 있다. 이러한 체계적 운영은 알고리즘 기반 서비스의 신뢰성을 강화하고, 잠재적 위험 요소를 사전에 예방하며, 지속적으로 개선하려는 노력을 반영한다.

[그림 4-20] (B)기업 자율검증 문항별 세부 점수



(B)기업의 자율검증 체계와 관련된 설문조사 역시 E1(자율검증 체계 구축), E2(체계 확인 및 개선), E3(체계 기록 및 보관), E4(체계 공개)로 구성되었으며, E1(자율검증 체계 구축)과 E2(체계 확인 및 개선) 문항에 모두 ‘예’로 응답하며 자율검증 체계 구축은 “자체적으로 기술적 오류, 이상 징후를 식별하고 있으며 AI와 휴먼 리뷰를 통해 자사의 커뮤니티 규정을 위반하는 콘텐츠와 계정을 검수하고 있는” 것으로 나타났다. 또한, 자율검증 체계를 지속적으로 관리하고 개선하려는 노력이 이루어지고 있음을 보여주었다. 반면, E3(체계 기록 및 보관)과 E4(체계 공개) 항목에는 “모른다”로 응답하며, 자율검증 체계의 기록과 보관 방식, 결과 공개 여부에 대한 인식이 명확하지 않은 것으로 나타났다. 이는 자율검증 체계와 관련된 정보 전달과 소통이 부족할 가능성을 시사한다. 전반적으로, 자율검증 체계 구축과 개선 면에서 강점을 보이며 알고리즘 기반 추천 서비스의 위험 관리를 체계적으로 실행하고 있는 것으로 평가된다. 그러나 자율검증 체계의 기록과 정보 공개 방식을 명확히 관리하고 운영하며 이를 외부에 투명하게 전달하는 부분에서 추가적인 개선이 필요하다는 점도 드러났다.

[그림 4-21] (A)기업과 (B)기업의 자율검증 문항별 세부 점수 비교



(A)기업과 (B)기업의 자율검증 실행 항목(E1E4)에 대한 비교 결과, E1(자율검증 체계 구축)과 E2(체계 확인 및 개선)에서는 두 플랫폼 모두 ‘예’로 응답하며 체계적인 검증 시스템을 운영하고 있음을 나타냈다. (A)기업은 E3(체계 기록 및 보관)과 E4(체계 공개) 포함 모든 항목에 대해 긍정적으로 응답하며 자율검증 체계를 철저히 관리하고 있음을 보여준다. 특히, 자율검증 체계가 추천 관련 정기 회의에서 검증되고 있으며 회사 내부의 누구나 회의록 열람이 가능하도록 하여 효과적으로 자율검증 체계를 마련하고 있음을 밝혔다.

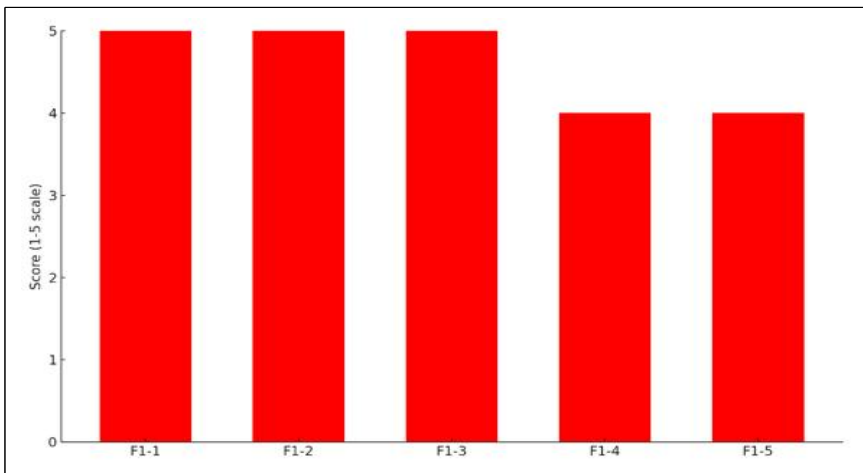
반면, (B)기업은 E1(자율검증 체계 구축)과 E2(체계 확인 및 개선)에서는 ‘예’로 응답하였으며, 기술적 오류와 이상 징후를 자체적으로 식별하고 AI와 휴먼 리뷰를 활용해 커뮤니티 규정을 위반하는 콘텐츠와 계정을 검토하는 검증 체계 운영의 구체적인 면모를 강조하였지만, E3(체계 기록 및 보관)과 E4(체계 공개)에 대해서는 응답하지 않아 기록 및 결과 공개와 관련된 구체적인 운영 방식이 명확히 드러나지 않았다.

전반적으로, (A)기업은 자율검증 체계를 전반적으로 철저히 이행하고 있으며 이를 기록하고 보관하며 공개하는 노력을 확인할 수 있다. (B)기업 역시 체계적인 검증 시스템을 운영하고 있으나 일부 항목에 대한 응답이 미흡하여 기록 관리와 결과 공개와 관련된 구체적인 개선이 필요함을 시사한다.

3) 불만 처리 및 분쟁 해결

F1(알고리즘 관련 불만 해결 노력), F1-1(고객대응 창구 안내), F1-2(연락 방법 안내의 명확성), F1-3(고객대응 창구 이용 가능 상황 안내), F1-4(스스로 문제를 해결할 수 있는 메뉴 제공), F1-5(다양한 고객대응 창구 운영), F1-6(고객대응 창구 운영 루트) 문항으로 이루어진 불만 처리 및 분쟁 해결 항목에 대한 (A)기업의 설문 결과에 따르면, 해당 기업은 이용자 보호 기본원칙 설문 문항 가운데 평균 4.6점으로 가장 높은 점수를 기록하며 알고리즘 추천 서비스와 관련된 궁금증이나 불만 사항, 분쟁을 해결하는 데 적극적으로 노력하고 있는 것으로 나타났다.

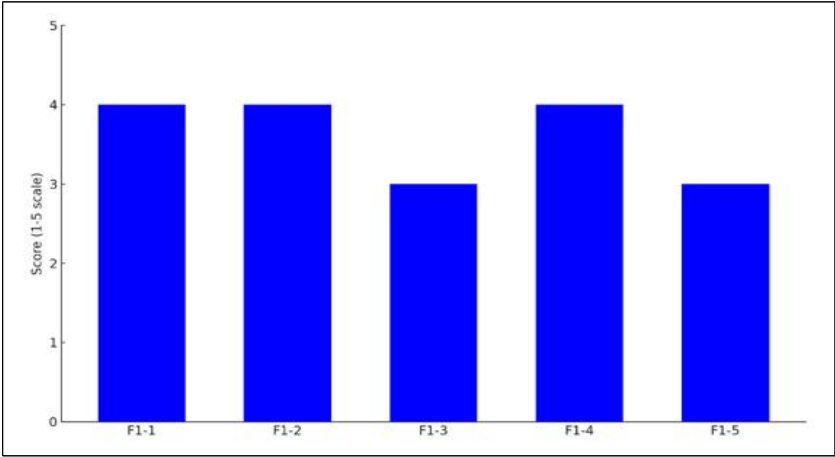
[그림 4-22] (A)기업 불만 처리 및 분쟁 해결 문항별 세부 점수



특히, F1(알고리즘 관련 불만 해결 노력) 항목에서 ‘예’로 응답하여 이용자 불만 해결을 위한 노력이 적극적으로 이루어지고 있음을 나타냈다. 또한, F1-1(고객대응 창구 안내), F1-2(연락 방법 안내의 명확성), F1-3(고객대응 창구 이용 가능 상황 안내) 문항에서는 5점(‘매우 그렇다’), F1-4(스스로 문제를 해결할 수 있는 메뉴 제공), F1-5(다양한 고객대응 창구 운영) 항목은 4점(‘그런 편이다’)의 높은 점수로 응답하며 (A)기업이 이용자에게 명확한 안내와 접근성을 제공하고 다양한 지원 루트를 운영하고 있음을 보여주었다. 특히, F1-6(고객대응 창구 운영 루트)에서는 “게시판, 채팅, 대표메일, 챗봇”이 제시되었으며 이를 통

해 (A)기업이 다양한 채널을 활용하여 이용자와의 소통과 문제 해결을 강화하고 있음을 확인할 수 있었다. 이러한 결과는 (A)기업이 불만 처리와 분쟁 해결에서 매우 적극적인 태도를 보이고 있음을 시사한다.

[그림 4-23] (B)기업 불만 처리 및 분쟁 해결 문항별 세부 점수

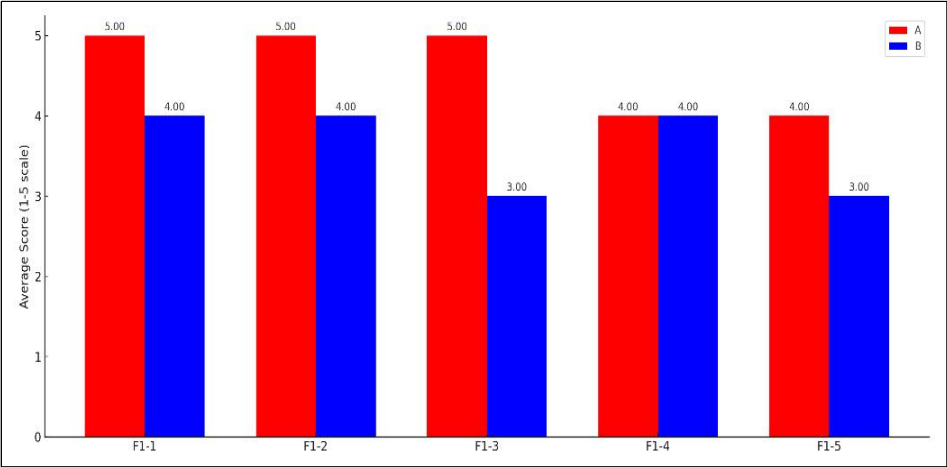


(B)기업의 불만 처리 및 분쟁 해결 항목에 대한 설문조사는 F1(알고리즘 관련 불만 해결 노력), F1-1(고객대응 창구 안내), F1-2(연락 방법 안내의 명확성), F1-3(고객대응 창구 이용 가능 상황 안내), F1-4(스스로 문제를 해결할 수 있는 메뉴 제공), F1-5(다양한 고객대응 창구 운영), F1-6(고객대응 창구 운영 루트) 문항으로 구성되었다. 분석 결과, F1(알고리즘 관련 불만 해결 노력)에서 ‘예’로 응답하며 이용자 불만 해결을 위한 체계를 충실히 운영하고 있음을 보여주었다. F1-1(고객대응 창구 안내), F1-2(연락 방법 안내의 명확성), F1-4(스스로 문제를 해결할 수 있는 메뉴 제공) 항목에서는 각각 4점(‘그런 편이다’)으로 응답되어 기본적인 문제 해결 체계가 마련되어 있음을 나타냈다. F1-3(고객대응 창구 이용 가능 상황 안내), F1-5(다양한 고객대응 창구 운영) 문항은 각각 3점(‘보통이다’)으로 응답되어 알고리즘 기반 추천 서비스와 관련된 불만 처리와 분쟁 해결에 있어 전반적으로 체계를 갖추고 노력하고 있으나 이용자에게 고객대응 창구를 쉽게 안내하거나 다양한 루트로 창구를 운영하는 측면에서는 개선이 필요한 상황임을 나타냈다. F1-5(다양한 고객대응

창구 운영)과 관련하여, F1-6(고객대응 창구 운영 루트)에서는 “신고 기능, 이메일, 자사 뉴스룸 등”이 활용되고 있다고 응답되어 (B)기업이 이용자와의 소통 강화를 위해 노력하고 있음을 보여주었다.

전반적으로, (B)기업은 불만 처리와 분쟁 해결에서 체계적이고 이용자 중심의 접근 방식을 채택하고 있으며 문제 해결을 위한 효율적인 체계를 운영하고 있는 것으로 평가되지만, 상대적으로 이용자가 구체적인 상황에서 어떤 창구를 활용할 수 있는지 명확히 알리고 고객대응 창구 운영 방식을 다양하게 운영하는 것이 주요 과제로 지적되며, 이 부분에서의 개선이 요구되는 상황이다.

[그림 4-24] (A)기업과 (B)기업의 불만처리 및 분쟁해결 문항별 세부 점수 비교



위 그래프는 (A)기업과 (B)기업의 불만 처리 및 분쟁 해결 항목(F1-1~F1-5)에 대한 비교를 보여준다. F1-1(고객대응 창구 안내)과 F1-2(연락 방법 안내의 명확성) 항목에서는 (A)기업이 각각 5점을 기록하며 (B)기업의 4점보다 높은 응답을 보였고, 접근성과 명확성 면에서 (A)기업이 더 강점을 보였다. F1-3(고객대응 창구 이용 가능 상황 안내)에서도 (A)기업이 5점을 기록하며 (B)기업의 3점보다 훨씬 높은 점수의 응답을 보여, 고객대응 창구 이용 가능 상황 안내에서 (A)기업의 우위를 나타냈다. F1-4(스스로 문제를 해결할 수 있는 메뉴 제공) 항목에서는 두 기업이 동일하게 4점을 기록하며, 문제 해결을 위한 자율적 접

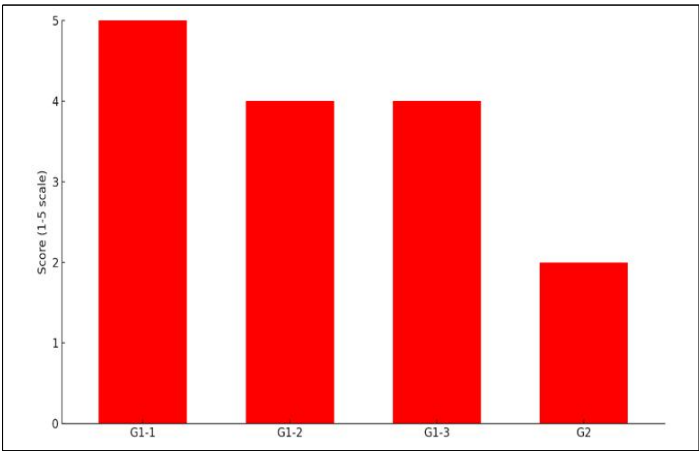
근 방식에서 유사한 수준의 노력을 보였다. F1-5(다양한 고객대응 창구 운영)에서는 (A)기업이 4점, (B)기업이 3점을 기록하며 고객 소통 채널의 다양성 측면에서 (A)기업이 더 높은 응답을 보였다.

주관식 응답에서는 (A)기업이 “게시판, 채팅, 대표메일, 챗봇”을, (B)기업이 “신고기능, 이메일, 자사 뉴스룸”을 주요 채널로 제시하며 접근 방식에서 다소 차이를 보였다. 전반적으로, (A)기업은 (B)기업에 비해 고객대응 창구 접근성과 명확성, 이용 가능 상황 안내 등에서 높은 응답을 보이며, 불만 처리와 분쟁 해결에서 비교적 강점을 보였다. (B)기업은 전반적으로 안정적인 수준의 응답을 기록했으나 고객대응 창구의 다양성과 정보 안내 측면에서 개선의 여지가 있음을 확인되었다.

4) 내부 규칙 제정

(A)기업의 내부 규칙 제정 측면에서 설문은 G1(내부 규칙 제정), G1-1(내부구성원 공유), G1-2(내부구성원 관련 교육), G1-3(고객대응 창구 안내), G1-4(고객대응 창구 운영 루트), G2(내부 규칙의 이용자 공개), G2-1(내부 규칙 공개 방식), 총 7개 문항으로 진행되었다.

[그림 4-25] (A)기업 내부 규칙 문항별 세부 점수

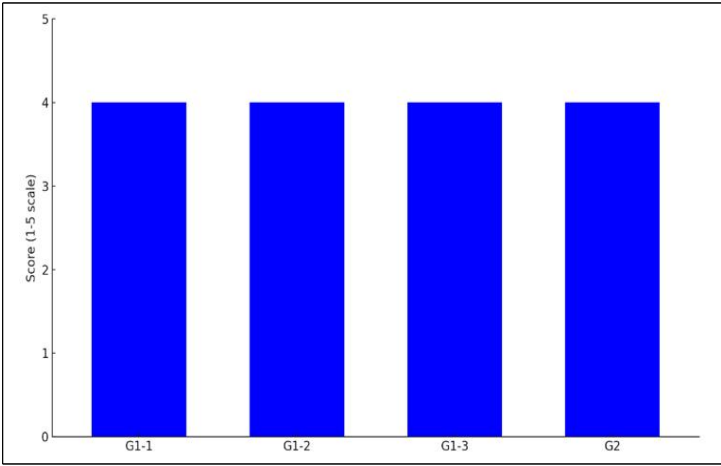


그 결과, (A)기업은 알고리즘 기반 추천 서비스의 운영 및 관리를 위한 내부 규칙을 마련하고 이를 내부 구성원들에게 공유하며 교육하는 데 적절한 노력을 기울이고 있는 것으

로 나타났다. G1(내부 규칙 제정) 항목에서 ‘예’로 응답하여 (A)기업이 내부 규칙을 자율적으로 제정하고 운영하고 있음을 보여주었다. G1-1(내부구성원 공유), G1-2(내부구성원 관련 교육), G1-3(고객대응 창구 안내) 항목에서도 모두 평균 4점(‘그런 편이다’) 이상의 점수로 응답하여 (A)기업이 내부 구성원과 효율적으로 소통하며 내부 규칙을 체계적으로 운영하고 있는 것으로 확인되었다.

그러나 G1-4(고객대응 창구 운영 루트)에 대한 응답은 이루어지지 않아 구체적인 정보가 확인되지 않았으며 G2(내부 규칙의 이용자 공개)와 G2-1(내부 규칙 공개 방식) 항목에서는 내부 규칙은 구체적이며 명시적으로 이용자들에게 공개되지 않는 편이며 “운영 및 관리에 대한 내부 규칙은 구체적으로 관리되고 있으나 이용자들에게는 공개되지 않고 있다”는 응답이 나왔다. 이를 통해 (A)기업이 내부적으로는 규칙을 적절히 관리하고 있지만, 외부 이용자와의 소통 및 정보 제공 측면에서는 부족한 점이 드러났다. 이는 해당 기업의 내부 규칙 운영의 투명성을 강화하고 이용자 신뢰를 확보하기 위해 일부 내부 규칙을 공개할 필요가 있음을 시사한다.

[그림 4-26] (B)기업 내부 규칙 문항별 세부 점수

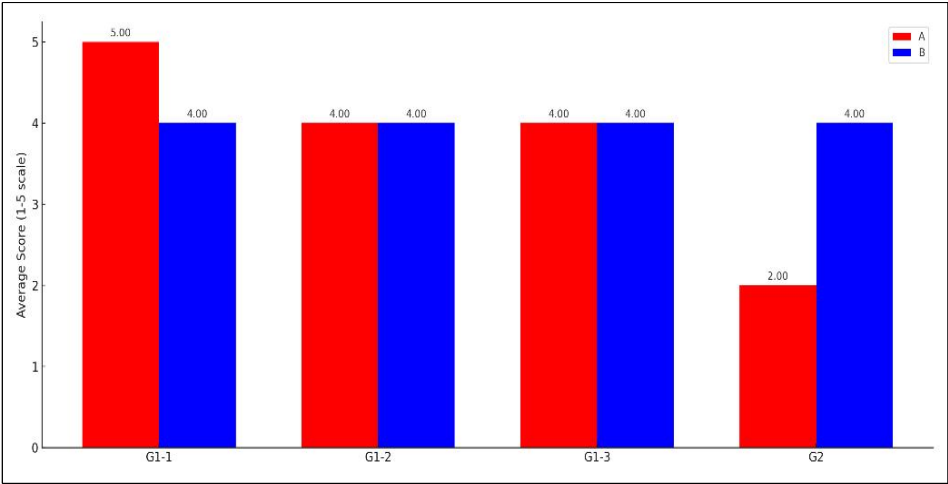


G1(내부 규칙 제정), G1-1(내부구성원 공유), G1-2(내부구성원 관련 교육), G1-3(고객대응 창구 안내), G1-4(고객대응 창구 운영 루트), G2(내부 규칙의 이용자 공개), G2-1(내부

규칙 공개 방식)로 구성된 (B)기업의 내부 규칙 제정 항목에 대한 설문조사 결과, G1(내부 규칙 제정)에서는 ‘예’로 응답되어, 내부 규칙을 자율적으로 제정하고 운영하고 있음을 보여주었다. G1-1(내부구성원 공유), G1-2(내부구성원 관련 교육), G1-3(고객대응 창구 안내), G2(내부 규칙의 이용자 공개) 항목은 모두 4점(‘그런 편이다’)의 응답을 기록하며, 해당 기업이 내부 규칙을 체계적으로 운영하고 명확한 안내를 제공하고 있음을 나타냈다. 한편, G1-4(고객대응 창구 운영 루트)는 ‘응답하지 않음’으로 기록되어 구체적인 정보는 확인되지 않았다. G2-1(내부 규칙 공개 방식)에 대한 응답에서는 “자사 Newsroom, Transparency Center 등”이 언급되며, 자사의 Newsroom과 Transparency Center를 활용해 내부 규칙을 구체적이며 명시적으로 이용자와 공유하고 있음을 보여주었다.

전반적으로, (B)기업은 알고리즘 추천 서비스의 운영 및 관리와 관련하여 내부 구성원을 대상으로 규칙을 마련하고 이를 공유하는 체계를 잘 갖추고 있다. 또한, 이용자에게도 내부 규칙을 비교적 적절히 공개하며, 내부와 외부 모두 고려한 균형 잡힌 접근 방식을 보여주고 있다. 이러한 내부 규칙의 체계적인 운영과 투명한 공개는 (B)기업이 이용자와의 신뢰를 강화하는 데 기여하고 있음을 나타낸다.

[그림 4-27] (A)기업과 (B)기업의 내부 규칙 문항별 세부 점수 비교



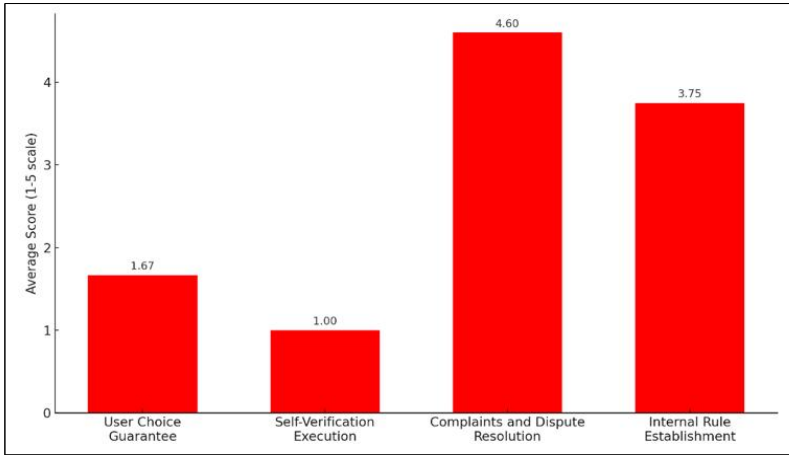
(A)기업과 (B)기업의 내부 규칙 제정 항목(G1-1~G1-3 및 G2)에 대한 비교 결과, G1-1(내부구성원 공유)에서는 (A)기업이 5점을, (B)기업이 4점을 기록하며 (A)기업이 내부 규칙을 내부구성원에게 명확히 공유하는 데 강점을 보였다. G1-2(내부구성원 관련 교육)와 G1-3(고객대응 창구 안내)에서는 두 기업이 모두 4점으로 응답하며 내부 규칙 관련 교육, 고객대응 체계에서 유사한 수준의 노력을 기울이고 있음을 나타냈다. G2(내부 규칙의 이용자 공개)에서는 (A)기업이 2점으로 낮은 응답을 기록한 반면, (B)기업은 4점으로 응답하며 내부 규칙 이용자 공개의 구체성과 명확성에서 (B)기업이 더 강점을 보였다. 하지만, G1-4(고객대응 창구 운영 루트)는 두 기업 모두 미응답하여 해당 항목에 대한 구체적인 정보는 확인되지 않았다.

주관식 응답에서는 (A)기업이 운영 규칙은 상세하지만 공개는 이루어지지 않다고 언급하며 내부 규칙의 대외적 투명성이 부족함을 드러낸 반면, (B)기업은 Newsroom과 Transparency Center에서 내부 규칙을 이용자에게 공개한다고 밝혀 접근성을 강조하였다. 전반적으로 (A)기업은 내부 규칙의 내부구성원 내 공유에서 강점을 보였으나, 외부 공개 측면에서는 (B)기업이 우위를 보이며 이용자 신뢰를 강화하기 위한 체계적인 접근을 보여주었다.

5) 실행 원칙 항목 비교 평가

실행 원칙 항목인 이용자 선택권 보장(User Choice Guarantee), 자율검증 실행(Self-Verification Execution), 불만 처리 및 분쟁 해결(Complaints and Dispute Resolution), 내부 규칙 제정(Internal Rule Establishment)에 대한 (A)기업의 평균 점수를 비교한 결과, 자율검증 실행 항목이 모든 응답에서 '예'로 표시되어 가장 높은 점수를 기록하며 해당 기업이 알고리즘의 잠재적 위험을 체계적으로 관리하고 검증하는 데 있어 우수한 노력을 기울이고 있음을 보여주었다.

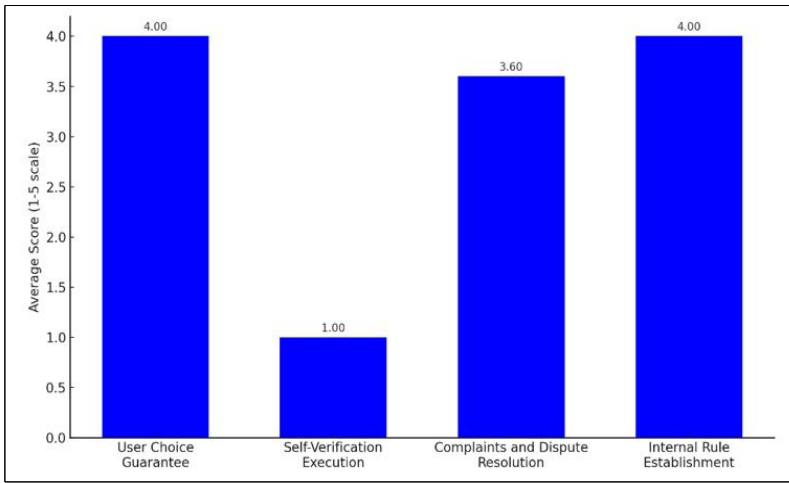
[그림 4-28] (A)기업 실행 원칙 문항별 평균 비교



불만 처리 및 분쟁 해결 항목과 내부 규칙 제정 항목도 각각 평균 4.6점, 3.75점으로 평균 4점에 가까운 높은 점수의 응답을 보여, 이용자의 불만 처리와 내부 규칙 운영 면에서 강점을 보이고 있는 것으로 나타났다. 반면, 이용자 선택권 보장 항목은 평균 1.67점으로 다른 항목에 비해 다소 낮은 점수를 기록하여 이용자에게 더 명확하고 구체적인 선택권 제공 방안이 필요한 개선 영역임을 시사했다. 전반적으로 현재 (A)기업은 자율검증과 불만 처리 및 분쟁 해결, 내부 규칙과 관련된 운영 체계에서 높은 성과를 보이고 있으나, 이용자 경험을 향상시키기 위해 선택권 보장 측면에서 추가적인 노력이 요구되는 상황이다.

(B)기업은 이용자 선택권 보장(User Choice Guarantee), 자율검증 실행(Self-Verification Execution), 불만 처리 및 분쟁 해결(Complaints and Dispute Resolution), 내부 규칙 제정(Internal Rule Establishment)에 대한 설문조사에서 전반적으로 높은 수준의 이행을 보인 것으로 나타났다.

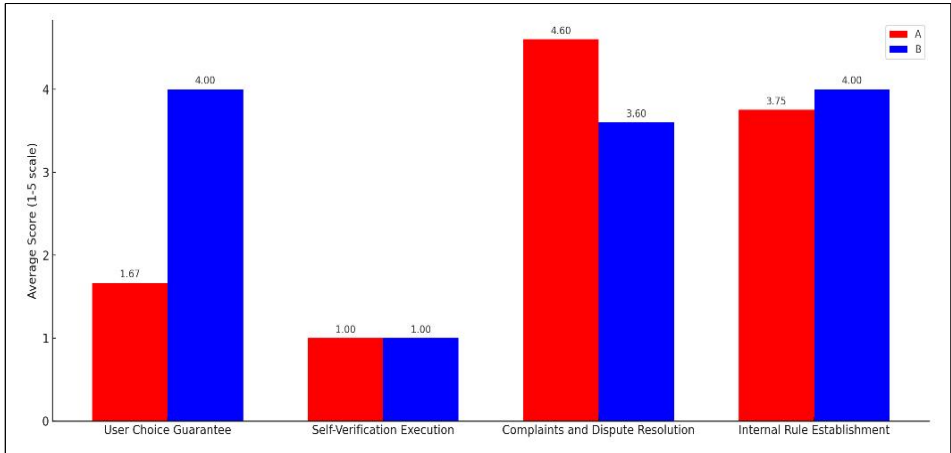
[그림 4-29] (B)기업 실행 원칙 문항별 평균 비교



자율검증 실행에서는 체계적인 검증 시스템을 운영하고 이를 지속적으로 관리하고 개선하려는 노력이 확인되었다. 기술적 오류와 이상 징후를 식별하며 규정을 위반하는 콘텐츠와 계정을 검토하는 등 내부적으로 검증 체계를 잘 구축하고 있지만, 기록 관리와 결과 공개 측면에서는 추가적인 개선이 필요함이 드러났다.

이용자 선택권 보장 항목은 평균 4점으로, 서비스 변경 가능성과 안내의 용이성에서 긍정적으로 평가하였으며, 내부 규칙 제정 항목 역시 평균 4점으로 명확한 규칙 운영과 이를 기반으로 한 이용자 신뢰 강화에서 강점을 보였다. 불만 처리 및 분쟁 해결 항목은 평균 3.6점으로, 다양한 고객대응 창구를 활용하며 이용자 친화적인 접근 방식을 보여주었으나 고객 창구 접근성과 구체화에 대한 필요성이 제기되었다. 전반적으로 (B)기업은 주요 항목에서 책임감 있는 운영과 이용자 중심의 접근 방식을 채택하며 높은 이행 수준을 보여주었지만 일부 세부적인 항목에서의 개선을 통해 더 높은 신뢰성과 완성도를 확보할 수 있을 것으로 보인다.

[그림 4-30] 실행 원칙 항목별 (A)기업과 (B)기업의 평균 비교



위 그래프는 두 기업의 실행 원칙(이용자 선택권 보장, 자율검증 실행, 불만 처리 및 분쟁 해결, 내부 규칙 제정)에 대한 비교를 보여준다. 자율검증 실행에서는 (A)기업이 모든 세부 항목에서 '예'로 응답하며 체계적이고 지속적인 검증 체계를 완전하게 운영하고 있음을 나타냈다. 반면, (B)기업은 두 개의 항목에서 '예'로 응답하며 검증 체계 구축과 관리·개선에 노력하고 있으나, 나머지 두 항목에 대해서는 응답하지 않아 일부 체계의 기록 및 결과 공개와 관련된 정보가 명확하지 않은 것으로 보인다.

내부 규칙 제정 항목에서는 (B)기업이 평균 4점, (A)기업이 3.75점을 기록하며 (B)기업이 더 체계적이고 명확한 규칙 운영과 관리를 보이고 있음을 나타냈다. (A)기업은 상대적으로 낮은 점수를 기록하여 내부 규칙의 정보 공개 측면에서 일부 개선이 필요한 것으로 보인다.

불만 처리 및 분쟁 해결 항목에서는 (A)기업이 평균 4.6점으로 (B)기업의 3.6점보다 높은 응답을 보이며 이용자 문제 해결과 분쟁 처리에서 더 안정적인 체계를 보유하고 있음을 나타냈다. 반면, (B)기업은 이 항목에서 비교적 낮은 응답을 보이며, 고객대응 창구 운영의 다양성과 접근성을 개선할 여지가 있는 것으로 나타났다.

이용자 선택권 보장 항목에서는 (B)기업이 평균 4점으로 높은 응답을 보인 반면, (A)기업은 1.67점으로 낮은 응답을 기록하며, 선택권 보장과 관련된 체계적 설계와 이용자 경험 개선이 절실히 요구됨을 시사했다. (B)기업은 선택권 제공과 관련된 체계를 비교적 잘 갖

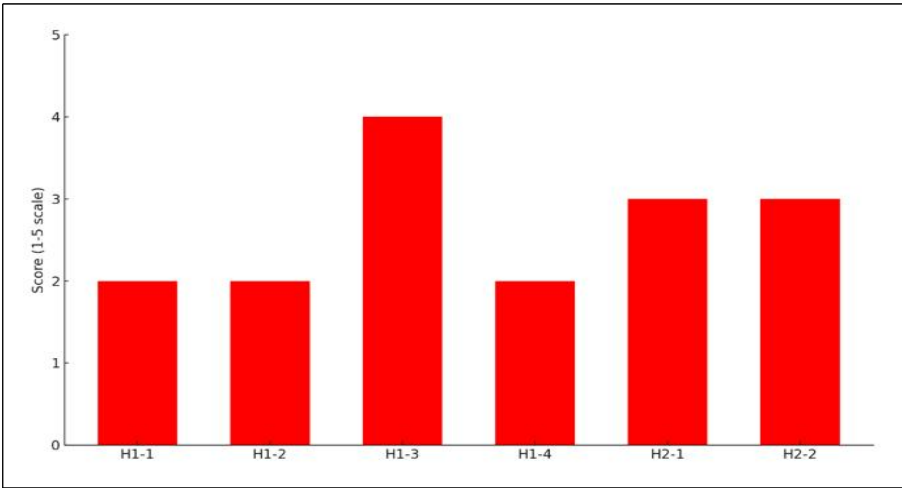
추고 있는 것으로 보인다.

전반적으로, (B)기업은 내부 규칙 제정과 이용자 선택권 보장에서 강점을 보이며, (A)기업은 불만 처리 및 분쟁 해결에서 더 긍정적인 응답을 기록하였지만, 두 기업 모두 특정 항목에서는 개선의 여지가 있음을 보여준다.

다. 정부 역할

알고리즘을 활용한 추천 서비스 이용자 보호와 관련한 정부의 역할과 관련하여 H1(정부 역할 - 서비스 제공자 지원) 항목은 H1-1(정부의 실행 가이드 제공), H1-2(지원 정책 수립), H1-3(자율규제 활동 지원), H1-4(기본원칙 준수 관리 감독)의 세부 문항으로, H2(정부 역할 - 이용자 역량 강화 및 보호) 항목은 H2-1(이용자의 인공지능 역량 강화 지원)과 H2-2(이용자 보호 책무 이행 촉구) 문항으로 구성되어 (A)기업을 대상으로 설문이 진행되었다.

[그림 4-31] (A)기업 정부 역할 문항별 세부 점수



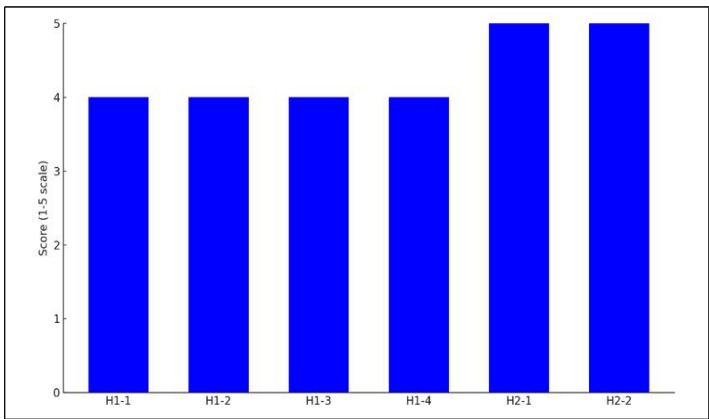
H1(정부 역할 - 서비스 제공자 지원) 항목에서는 H1-3(자율규제 활동 지원)이 4점(‘약간 동의한다’)으로 응답된 것을 제외하고, H1-1(정부의 실행 가이드 제공), H1-2(지원 정책 수립), H1-4(기본원칙 준수 관리 감독) 항목이 모두 2점(‘별로 동의하지 않는다’)으로 응답되

었다. 이를 통해 (A)기업이 정부의 개입에 대해 전반적으로 부정적인 태도를 보이며, 자율 규제를 중시하는 입장을 취하고 있음을 확인하였다. 기본원칙에 대한 실행 가이드 제공, 지원 정책 수립, 기본원칙 준수 여부를 관리 감독하는 정부 역할에 대해서는 부정적인 반응을 나타냈으며 H1-5(정부 역할 관련 추가 의견)에 대해서는 응답이 이루어지지 않아 구체적인 추가 의견은 확인되지 않았다.

H2(정부 역할 - 이용자 역량 강화 및 보호) 항목에서는 H2-1(이용자의 인공지능 역량 강화 지원)과 H2-2(이용자 보호 책무 이행 촉구)가 모두 3점(‘보통이다’)으로 응답하여 (A)기업이 정부의 이용자 보호 및 역량 강화 역할에 대해 어느 정도 동의하지만, 이를 우선순위로 두지는 않고 있음을 보여주었다. H2-3(정부 역할 관련 추가 의견) 역시 응답하지 않아 구체적인 추가 의견은 확인되지 않았다.

H1(서비스 제공자 지원)과 H2(이용자 역량 강화 및 보호) 항목 중 서비스 제공자 지원 측면에서는 평균 2.5점, 이용자 역량 강화 및 보호 부분에서는 평균 3점 점수를 기록하여, (A)기업이 정부의 두 가지 역할에 공통적으로 다소 부정적인 태도를 가지고 있음을 보여 준다. H1은 서비스 제공자 지원을 중심으로 낮은 평가를 받았고 H2는 이용자 보호와 역량 강화를 중심으로 서비스 제공자 지원 항목 대비 상대적으로 높은 평가를 받았지만 두 항목 모두 추가 의견 제공이 부족하여 구체적인 협력 방안 마련에는 더 많은 논의가 필요함을 시사한다.

[그림 4-32] (B)기업 정부 역할 문항별 세부 점수



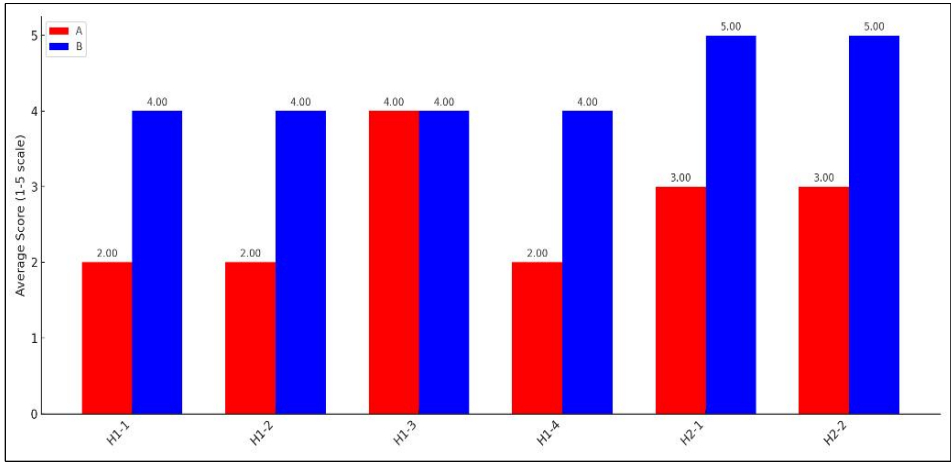
(B)기업도 동일하게 H1(정부 역할 - 서비스 제공자 지원) 항목은 H1-1(정부의 실행 가이드 제공), H1-2(지원 정책 수립), H1-3(자율규제 활동 지원), H1-4(기본원칙 준수 관리 감독)의 세부 문항으로, H2(정부 역할 - 이용자 역량 강화 및 보호) 항목은 H2-1(이용자의 인공지능 역량 강화 지원)과 H2-2(이용자 보호 책무 이행 촉구)로 구성되어 설문이 진행되었다.

H1(정부 역할 - 서비스 제공자 지원)에 해당하는 H1-1(정부의 실행 가이드 제공), H1-2(지원 정책 수립), H1-3(자율규제 활동 지원), H1-4(기본원칙 준수 관리 감독)은 모두 4점으로 응답되며 (B)기업이 정부의 서비스 제공자 지원 역할에 대해 긍정적으로 평가하고 있음을 보여준다. 이는 해당 기업이 정부의 정책적 지원과 자율규제 활동에 대한 협력을 중요하게 생각하고 있음을 시사한다. 한편, H1-5(정부 역할 관련 추가 의견)은 응답하지 않아 구체적인 추가 의견은 확인되지 않았다.

H2(정부 역할 - 이용자 역량 강화 및 보호)에 해당하는 H2-1(이용자의 인공지능 역량 강화 지원)과 H2-2(이용자 보호 책무 이행 촉구)는 모두 5점으로 응답되어 (B)기업이 정부의 이용자 보호와 역량 강화 역할에 대해 매우 긍정적으로 평가하고 있음을 보여주며 이용자 중심의 정부 역할에 대해 높은 기대와 신뢰를 가지고 있음을 나타낸다. 한편, H2-3(정부 역할 관련 추가 의견)에 대하여 미응답하여 구체적인 추가 의견은 확인되지 않았다.

H1(서비스 제공자 지원) 항목은 평균 4점, H2(이용자 역량 강화 및 보호) 항목은 평균 5점으로 응답되었다는 점은 (B)기업이 정부의 양쪽 역할에 대해 신뢰를 가지고 긍정적으로 평가하고 있음을 보여준다. H1은 서비스 제공자 지원 측면에서 안정적인 응답을, H2는 이용자 보호와 역량 강화 측면에서 매우 긍정적인 응답을 보였다. 그러나 두 항목 모두에서 추가 의견 제공이 부족하여 구체적인 협력 방안 마련에는 더 많은 논의가 필요함을 시사한다.

[그림 4-33] (A)기업과 (B)기업 정부 역할 문항별 세부 점수 비교

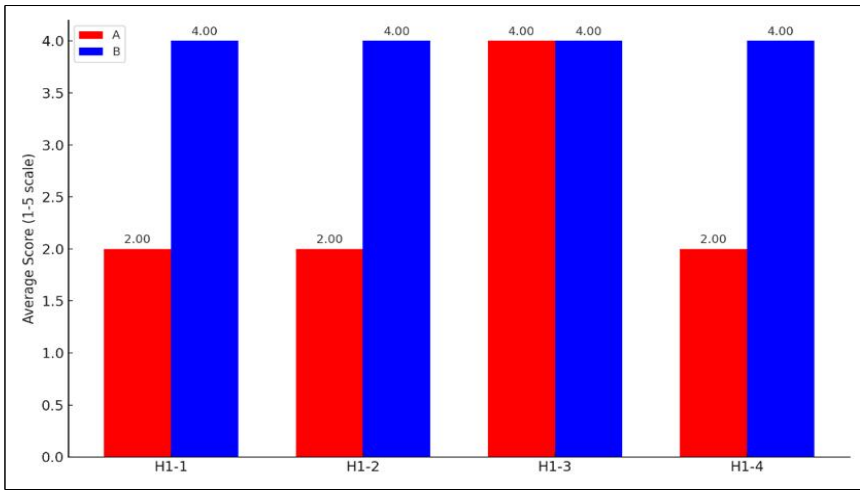


위 그래프는 (A)기업과 (B)기업의 정부 역할에 대한 응답을 비교한 것이다. H1-1(정부의 실행 가이드 제공)과 H1-2(지원 정책 수립)에서 (A)기업은 각각 2점으로 응답한 반면, (B)기업은 4점으로 응답하며 (B)기업이 정부의 서비스 제공자 지원 역할에 대해 더 긍정적인 견해를 가지고 있음을 보여준다. H1-3(자율규제 활동 지원)에서는 두 기업 모두 4점으로 응답하며 정부의 자율규제 활동 지원에 대해서는 비슷한 수준의 긍정적인 인식을 보였다. H1-4(기본원칙 준수 관리 감독)에서는 (A)기업이 2점, (B)기업이 4점으로 응답하여 (B)기업이 정부의 기본원칙 준수 관리 감독 역할에 대해 더 긍정적인 시각을 가지고 있음을 나타냈다.

H2-1(이용자의 인공지능 역량 강화 지원)과 H2-2(이용자 보호 책무 이행 촉구)에서는 (A)기업이 각각 3점으로 응답한 반면, (B)기업은 각각 5점으로 응답하여, (B)기업이 정부의 이용자 역량 강화 및 보호 역할에 대해 보다 더 긍정적인 견해를 보이고 있음을 알 수 있다.

전반적으로, (A)기업은 정부의 서비스 제공자 지원 및 이용자 보호 역할에 대해 다소 제한적인 인식을 보였으나 자율규제 활동 지원에는 비교적 긍정적으로 응답하였다. 반면, (B)기업은 전반적으로 정부의 역할에 대해 긍정적인 견해를 보이며 특히 이용자 보호와 역량 강화에 있어 높은 기대와 신뢰를 나타냈다.

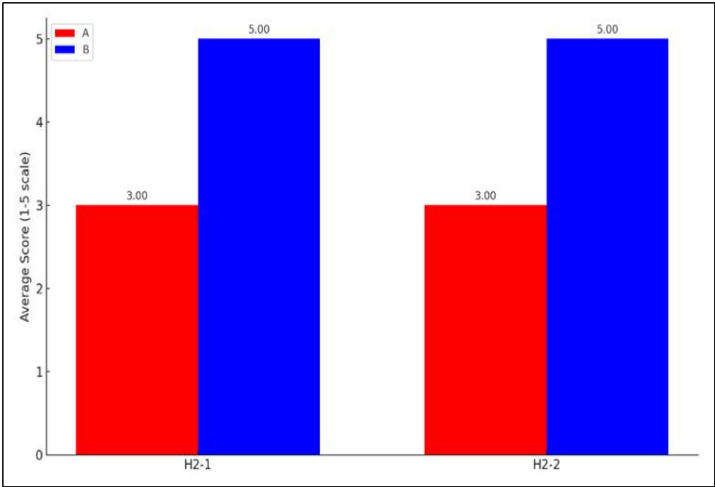
[그림 4-34] (A)기업과 (B)기업 서비스 제공자 지원(H1) 문항별 세부 점수 비교



보다 구체적으로, (A)기업과 (B)기업의 정부 역할 중 서비스 제공자 지원(H1) 항목에 대한 견해를 비교한 결과, H1-1(정부의 실행 가이드 제공)과 H1-2(지원 정책 수립)에서 (A)기업은 각각 2점으로 응답한 반면, (B)기업은 4점으로 응답하며 (B)기업이 정부의 서비스 제공자 지원 역할에 대해 더 긍정적인 인식을 가지고 있음을 나타냈다. 이는 (B)기업이 정부의 정책적 지원과 실행 가이드 제공에 대해 높은 기대와 신뢰를 보이고 있는 반면, (A)기업은 이 부분에서 다소 제한적인 견해를 보였음을 시사한다. H1-3(자율규제 활동 지원)에서는 두 기업이 모두 4점으로 응답하며 정부의 자율규제 활동 지원 역할에 대해 비슷한 수준의 긍정적인 시각을 가지고 있음을 보여준다. 이 항목은 양측이 모두 중요하게 인식하고 있는 부분으로, 정부의 역할이 자율적인 규제 활동을 뒷받침할 수 있는 환경 조성에 기여해야 한다는 공통된 인식을 엿볼 수 있다. H1-4(기본원칙 준수 관리 감독)에서는 (A)기업이 2점, (B)기업이 4점으로 응답하여 (B)기업이 정부의 기본원칙 준수와 관련된 관리 및 감독 역할에 대해 더 긍정적인 견해를 보였다. 이는 (B)기업이 기본원칙의 실행과 관리 감독에 있어 정부의 적극적인 개입을 더 신뢰하고 있다는 점을 보여준다. 반면, (A)기업은 이와 관련해 다소 제한적인 인식을 보이며 정부의 관리 감독 역할에 대한 기대가 상대적으로 낮은 것으로 나타났다.

전반적으로 (B)기업은 모든 항목에서 더 높은 응답을 기록하며 정부의 서비스 제공자 지원 역할에 대해 체계적이고 긍정적인 견해를 보였다. 특히, 실행 가이드 제공과 지원 정책 수립, 기본원칙 준수 관리 감독과 같은 항목에서 (A)기업과의 점수 차이가 두드러지며 (B)기업이 정부의 정책적 지원과 개입을 중요한 요소로 인식하고 있음을 확인할 수 있다. (A)기업은 자율규제 활동 지원에 대해서는 긍정적인 견해를 보였으나 기타 항목에서는 정부의 지원과 역할에 대해 상대적으로 낮은 기대를 보였다. 이는 두 기업의 정부 역할에 대한 관점과 기대의 차이를 잘 보여주는 결과라고 할 수 있다.

[그림 4-35] (A)기업과 (B)기업 이용자 역량 강화 및 보호(H2) 문항별 세부 점수 비교

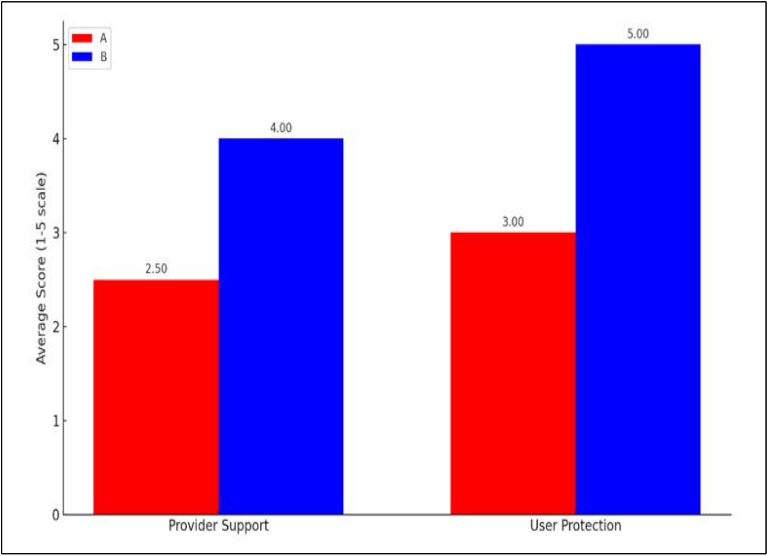


또 다른 그래프는 (A)기업과 (B)기업의 정부 역할 중 이용자 역량 강화 및 보호(H2) 항목에 대한 견해를 비교한 결과를 보여준다. H2-1(이용자의 인공지능 역량 강화 지원)에서 (A)기업은 3점, (B)기업은 5점으로 응답하며 (B)기업이 정부의 인공지능 관련 이용자 역량 강화 지원 역할에 대해 매우 긍정적인 견해를 보인 반면, (A)기업은 상대적으로 낮은 기대를 드러냈다. 이는 (B)기업이 정부와 협력하여 이용자가 인공지능 기술을 이해하고 활용할 수 있는 환경을 조성하는 데 더 적극적인 역할을 강조하고 있음을 시사한다. 반면, (A)기업은 이러한 지원 필요성에 대해 다소 제한적인 인식을 보이고 있다. H2-2(이용자 보호 책무

이행 촉구)에서도 동일하게 (A)기업이 3점, (B)기업이 5점으로 응답하며 (B)기업이 정부의 이용자 보호 역할에 대해 더 높은 신뢰와 기대를 보였다. (B)기업은 정부가 이용자의 권리를 보호하고 책임감을 가진 기업 운영을 촉구하는 데 있어 강력한 역할을 수행해야 한다고 인식하는 반면, (A)기업은 이 부분에서도 정부 지원의 필요성에 어느 정도 동의하지만 이를 우선순위로 두지는 않고 있다는 인식을 보이고 있다.

이러한 결과는 (B)기업이 정부와의 협력을 통해 이용자 보호와 역량 강화를 중요한 과제로 삼고 있음을 보여준다. 특히, (B)기업은 정부의 정책과 지원을 통해 이용자들이 더 나은 기술적 역량을 갖추고 기업의 책임 있는 운영을 보장받을 수 있는 환경을 만드는 데 적극적인 입장을 취하고 있다. 반면, (A)기업은 이용자 보호와 역량 강화에 대한 정부의 역할에 대해서도 마찬가지로 다소 제한적인 기대를 보이며 이는 기업의 운영 방식이나 이용자 지원 정책에서 정부 역할에 대한 관점 차이를 잘 보여주는 결과로 볼 수 있다.

[그림 4-36] 정부 역할 항목별 (A)기업과 (B)기업의 평균 비교



서비스 제공자 지원(Provider Support)과 이용자 역량 강화 및 보호(User Protection) 항목으로 구성된 정부 역할에 대한 (A)기업과 (B)기업의 견해를 비교한 결과, 서비스 제공자

지원 항목에서는 (A)기업이 평균 2.5점, (B)기업이 평균 4점으로 응답하며 (B)기업이 정부의 실행 가이드 제공, 지원 정책 수립, 자율규제 활동 지원 등 서비스 제공자 지원 역할에 대해 보다 긍정적인 견해를 보이고 있음이 나타났다. 이는 (B)기업이 정부의 정책적 지원과 개입을 중요한 요소로 인식하는 반면 (A)기업은 이 부분에서 다소 제한적인 기대를 가지고 있음을 시사한다.

이용자 역량 강화 및 보호 항목에서는 (A)기업이 평균 3점, (B)기업이 평균 5점으로 응답하며 (B)기업이 정부의 이용자 보호와 역량 강화 역할에 대해 매우 긍정적인 견해를 드러냈다. 특히, (B)기업은 정부가 인공지능 기술 이해와 활용 능력을 강화하는 데 있어 중요한 역할을 해야 한다고 보고 있으며, 이용자 권리 보호를 위한 기업 운영을 촉구하는 정부의 역할을 높이 평가하고 있음을 알 수 있다. 반면, (A)기업은 이용자 보호와 역량 강화에서 정부의 역할에 대해 보다 보수적인 입장을 보여 이 부분에서 기업 간의 인식 차이가 드러났다.

전반적으로, (B)기업은 서비스 제공자 지원과 이용자 보호 및 역량 강화에 대한 정부의 역할에 대해 매우 긍정적인 견해를 보이며, 정부와의 협력을 통해 이를 실현하려는 강한 의지를 나타냈다. 반면, (A)기업은 정부의 역할에 대해 다소 제한적인 기대를 보이며 정부와의 협력과 지원에 있어 비교적 소극적인 입장을 취하는 것으로 보인다. 이러한 결과는 두 기업의 정부 역할에 대한 인식 차이를 명확히 보여주고 있다.

4. 심층 인터뷰

가. 심층 인터뷰 개요

본 연구의 인터뷰는 설문조사를 통해 수집된 응답을 기반으로 추가적인 정보를 확인하고 심층적인 의견을 수렴하기 위해 실시되었다. 설문조사에서 도출된 주요 응답 중 구체적인 사례나 세부적인 설명이 필요한 문항에 대해 각 사업자를 대상으로 인터뷰를 진행하여 추가적인 정보를 청취하였다. 이러한 인터뷰 과정을 통해 설문 응답의 맥락을 보완하고 사업자의 실제 정책 및 운영 상황에 대한 보다 깊은 이해를 도모하고자 하였다.

(A)기업을 대상으로 한 인터뷰는 설문조사 문항 및 응답을 기본으로 진행되었으며, 이를 통해 도출된 주요 응답 중 구체적인 사례나 세부적인 설명이 필요한 문항에 대해 심층적

으로 논의하였다. 추가적으로, 설문조사에서 응답한 내용을 심화하고 기본원칙의 실효성을 더욱 구체화하기 위한 질문이 이루어졌다.

투명성과 관련하여, AI 추천 서비스 제공 사실을 이용자에게 실시간 문자나 홈화면 알림 등의 방식으로 통지하는 현행 방식의 효과를 묻고, OTT 서비스의 특성을 고려했을 때 더 적합한 대안이 있는지 논의하였다. 추천 알고리즘의 설명 측면에서 콘텐츠 자동 배열의 주요 기준(예: 시청 이력, 조회 수 등)을 이용자에게 설명할 때, 이용자들이 가장 이해하기 쉬운 방식에 대한 의견도 청취하였다. 이용자 선택권 보장과 관련해서는 추천 알고리즘의 매개변수(예: 시청 이력 반영 비중, 장르 선호도 등)를 이용자가 조정할 수 있도록 하는 기술적 가능성과 바람직한 허용 수준에 대해 질문하였다. 자율검증 체계의 실효성과 관련해서는 추천 시스템의 편향성이나 문제점 발견 시 이를 수정하는 데 소요되는 시간과 비용에 대해 구체적으로 논의하였다.

이와 함께, 추천 시스템 관련 이용자 불만 처리 메커니즘에서는 가장 빈번한 불만 유형과 이를 해결하기 위해 기본원칙에 포함될 추가적인 내용에 대한 의견을 들었으며, 연령별 보호조치와 관련하여 미성년자 등 특정 연령대 이용자를 위한 맞춤형 추천 시스템 보호조치의 필요성과 구현 가능한 보호장치에 대해서도 논의가 이루어졌다.

AI 기술 발전에 대응하는 방안으로, 새로운 추천 기술의 도입을 저해하지 않으면서도 이용자 보호를 달성하기 위해 기본원칙이 어떤 방식으로 개선되어야 하는지 물었고, 내부 규칙 제정의 실효성을 높이기 위해 필요한 구체적인 가이드라인에 대한 의견도 청취하였다. 마지막으로, 기본원칙 준수를 위해 필요한 정부의 지원책과, 국내 OTT 사업자들이 이를 원활히 이행하기 위해 어떤 지원이 필요한지, 그리고 글로벌 OTT 서비스들의 추천 시스템 관련 이용자 보호 정책 중 국내 도입이 필요하다고 생각되는 우수 사례에 대한 의견을 수렴하였다.

반면, (C)기업 인터뷰에서는 글로벌 OTT 사업자로서 알고리즘 추천 서비스 운영에 대한 구체적인 사례와 지역별 차별화 전략을 확인하고 한국 시장에 특화된 접근 방식을 이해하기 위해 다양한 질문이 사용되었다. 먼저, 글로벌-로컬 전략에서 투명성 원칙이 어떻게 적용되는지를 파악하기 위해 전 세계적으로 서비스되는 콘텐츠 추천 알고리즘의 주요 요소들(예: 시청 기록, 시청 완료율, 장르 선호도 등)이 국가별 또는 지역별로 다르게 가중치를 부여받는지, 그리고 각 국가의 규제 정책이나 문화적 특성에 따라 이러한 요소들의 투명

성 공개 수준을 조정하는 방식에 대해 논의하였다. 또한, 전 세계 이용자들에게 일관된 사용자 경험을 제공하면서도, 국가별로 상이한 알고리즘 통제 권한을 이용자에게 제공하는 방식을 어떻게 조화시키는지 질문하였다. 특히 한국 시장을 중심으로, 어느 수준까지의 이용자 통제권이 적절하다고 생각하는지를 묻고, 글로벌 서비스에서 이용자 선택권 보장의 범위를 탐구하였다.

(C)기업이 자체 제작 콘텐츠(오리지널 시리즈 등)와 라이선스 콘텐츠에 대해 추천 알고리즘을 어떻게 적용하고 있는지도 확인하였다. 특정 프랜차이즈나 장르가 과도하게 추천되지 않도록 내부적으로 어떤 기준을 설정하고 있는지 질문하며, 콘텐츠 포트폴리오 특성을 반영한 공정성 기준을 살펴보았다. 또한, 다양한 국가의 규제 환경 속에서 자율검증 체계를 운영할 때 가장 중점을 두는 부분과 함께, 한국의 자율검증 가이드라인에 대해 제안할 만한 개선사항이 있는지도 물었다. 이를 통해 글로벌 스탠다드와 한국 시장의 규제 요구사항 간의 정합성을 검토하고자 하였다.

글로벌-로컬 관점에서의 불만 처리 체계를 이해하는 데 초점을 맞춰, 각 국가별로 추천 알고리즘과 관련한 이용자 불만의 유형과 빈도가 다른지를 파악하고, 글로벌 차원의 표준화된 불만 처리 프로세스와 국가별 특화된 대응 방식이 어떻게 조화를 이루는지 확인하였다. 더 나아가, 유럽연합(EU)의 DSA(Digital Services Act)나 GDPR(General Data Protection Regulation)과 같은 강력한 디지털 규제를 준수하면서 축적된 경험이 한국의 알고리즘 추천 서비스 투명성 제고에 어떻게 기여할 수 있는지, 특히 이용자 통지 방식과 관련해 제안할 만한 모범 사례가 있는지 질문하였다. 이를 통해 글로벌 규제 대응 경험의 한국 시장 적용 가능성을 모색하고자 했다.

마지막으로, 글로벌 OTT 시장에서 한국이 가지는 특수성(예: 높은 디지털 리터러시, K-콘텐츠의 국제적 영향력 등)을 고려했을 때, 알고리즘 추천 서비스의 투명성과 공정성을 강화하기 위한 기본원칙이 어떤 방향성을 가져야 한다고 보는지 논의하였다. 이를 통해 한국 시장에서 글로벌 사업자가 가지는 관점을 이해하고, 추천 서비스의 원칙 설정에 대한 시사점을 도출하였다.

나. 심층 인터뷰 결과

1) 투명성

인공지능 기반 미디어 추천 서비스 이용자 보호 기본원칙 가운데 핵심 원칙인 투명성과 관련하여 (A)기업은 알고리즘 기반 추천 서비스의 구체적인 종류를 공개하지 않는 이유에 대해 구체적인 입장을 밝혔다. (A)기업은 현재 관련 가이드라인과 강제성이 명확하지 않으며, 알고리즘 공개는 사업자의 고유 권한을 노출하는 측면이 있어 이를 꺼린다고 설명하였다. 현재 서비스에서 배너의 알고리즘 추천 여부는 명시적으로 표시되지 않고 있으며, 추천 콘텐츠에 대한 알림 서비스는 제공되지만, 해당 알림 역시 알고리즘에 기반한 것임을 구체적으로 안내하지 않고 있다. (A)기업의 서비스 운영 방식은 알고리즘보다는 큐레이션 중심으로 구성되어 있으며, 밴드 순서의 경우, 일부만 개인화되고 있다. 반면, 콘텐츠는 주로 시청 기록을 기반으로 개인화 추천되지만 성별이나 연령과 같은 개인정보는 활용하지 않고 있다고 강조하였다. 개인정보 처리방침에서 개인화 요소를 간략히 언급하고 있어 이용자들이 개인화 알고리즘의 작동 방식을 명확히 인지하지는 못하는 상황이지만, (A)기업은 별도로 명시하지 않더라도 이용자들이 밴드가 개인화 추천임을 어느 정도 인지하고 있다는 입장을 전했다. 추천 알고리즘의 설명 필요성에 대해 (A)기업은 현재 고객들이 추천 알고리즘에 큰 관심을 보이지 않고 있으며, 관련 문의는 거의 없는 상황이라고 밝혔다. 추가적인 문의 사항이 발생할 경우 고객센터에서 개별 응대 가능함을 덧붙이며 이러한 시스템이 현재로서는 적합하다고 판단하고 있음을 나타냈다. (A)기업은 추천 알고리즘 서비스에 대한 구체적인 이용자 정보 제공을 위해서는 이용자 보호 측면 등 명확한 목적성이 필요하다는 의견을 제시하였다.

(C)기업은 알고리즘보다는 인간 중심의 콘텐츠 편성 전략을 활용하고 있음을 설명했다. (C)기업은 지역별 선호도와 메타데이터 분석을 기반으로 콘텐츠를 추천하고 있으며, 각 지역에서 중요하다고 판단되는 콘텐츠를 우선 배치한다고 밝혔다. 또한, 지역적 선호도를 반영하여 콘텐츠 계약과 편성이 이루어지고 있어 국가별로 콘텐츠 노출이 달라지는 경우도 발생한다고 전했다. 따라서, (C)기업은 기술적 투명성보다는 지역적 맥락과 선호도에 따른 콘텐츠 구성의 설명이 중요한 요소로 작용하고 있다. 알고리즘 기반의 자동화된 추천보다는 인간의 판단과 수동적 전략이 주요 방식으로 활용되고 있다.

두 기업의 접근 방식은 추천 서비스 설계와 운영에서 뚜렷한 차이를 보여준다. (A)기업

은 알고리즘 기반 추천을 중심으로 서비스를 운영하며 시청 기록을 주요 기준으로 개인화 추천을 제공한다. 그러나 성별, 연령 등 민감한 개인정보는 활용하지 않으며 개인정보 처리방침에 개인화 요소를 간략히 언급하는 데 그치고 있다. 베너나 알림에서 알고리즘 추천 여부를 명시적으로 표시하지 않고 있으며 큐레이션 중심의 설계와 일부 개인화된 밴드 순서를 활용해 서비스를 구성하고 있다. 반면, 이용자 문의가 거의 없다는 점을 근거로 알고리즘에 대한 구체적인 설명 필요성을 느끼지 않는다는 입장을 밝혔다. 추가적인 정보 제공은 이용자 보호와 같은 명확한 목적성이 제기될 경우에 검토할 수 있다고 덧붙였다. (C)기업은 알고리즘보다는 인간 중심의 콘텐츠 편성 전략을 채택하고 있으며 지역별 선호도와 메타데이터 분석에 기반해 콘텐츠를 추천하고 배치한다. 특정 지역의 문화적 맥락과 선호도를 반영해 콘텐츠 계약과 편성을 결정하며 이는 국가별로 콘텐츠 노출이 다르게 나타나는 결과를 초래한다. 기술적 투명성을 강조하기보다는 지역적 특성에 따른 콘텐츠 구성을 우선시하며, 알고리즘 기반의 자동화보다는 인간 판단과 수동적 전략을 중요하게 활용한다는 점에서 차이를 보인다. 결과적으로, (A)기업은 알고리즘 중심의 개인화된 추천 서비스에 중점을 두면서도 이용자 보호와 투명성을 보완할 여지를 남겨둔 신중한 접근을 보이고 있다. 반면, (C)기업은 지역별 선호도를 기반으로 한 콘텐츠 편성 전략을 통해 서비스를 제공하며 인간의 판단을 주요 수단으로 활용하고 있다. 두 기업은 각각의 목표와 서비스 특성에 따라 기술 기반의 접근과 인간 중심의 접근이라는 상이한 전략을 채택하고 있다.

2) 공정성

공정성 확보 측면에서 (A)기업은 추천 알고리즘에서 특정 채널이나 콘텐츠에 가중치를 부여하거나 차별적으로 배치하지 않는다고 밝혔다. 클릭 빈도와 데이터 분석을 통해 추천 모델을 정교화하고 있으며, 어뷰징 방지를 위해 이상 클릭 데이터를 검수하여 랭킹에서 제외하는 절차를 운영하고 있다. 현재 오리지널 콘텐츠 등 특정 채널을 기준으로 가중치를 부여하고 있지 않으며, 편성 운영에서도 수동적인 개입이 거의 없는 점을 강조하였다. 팝업이나 배너를 통해 특정 콘텐츠를 홍보하는 경우에도, 이는 단순한 프로모션 활동으로 알고리즘에 직접적인 영향을 미치지 않는다고 설명했다. ‘공정성’의 정의와 관련하여 (A)기업은 내부 및 외부의 이해관계자에게 특혜를 배제하고 이용자에게는 ‘정확성’을 중심으로 서비스를 제공하는 것이 공정성을 실현하는 핵심이라는 의견을 밝혔다. 장르의 다양성을

자연스러운 노출을 통해 확보하려는 노력을 기울이고 있으며, 내부적으로 매주 정기 회의가 진행되어 시스템 개선과 기술적 고도화를 논의하고 있다. 구체적으로는 기술적 고도화를 위해 BERT 모델을 활용한 유사 콘텐츠 추천 기술과 하이퍼파라미터 조정을 통해 단기 및 장기 시청 패턴을 분석할 수 있는 시스템을 도입하고 있다.

(C)기업은 공정성 원칙을 준수하는 데 있어 오리지널 콘텐츠와 라이선스 콘텐츠 간에 추천 알고리즘의 차이를 두지 않고 있다고 밝혔다. 모든 콘텐츠의 추천은 비즈니스적 목적, 즉 이용자 확보에 초점이 맞추어져 있다. 따라서 콘텐츠 추천과 관련된 공정성 측면에서 추천 알고리즘이 콘텐츠를 배치하는 방식은 사용자 선호와 데이터에 기반하며, 비즈니스적 전략이 우선시된다고 전했다. 특히, 투입된 마케팅 예산의 차이가 콘텐츠 노출 빈도에 영향을 미치는 것은 알고리즘의 공정성보다는 이용자 선호를 기반으로 콘텐츠를 배치하려는 의도와 연관된다고 설명하였다.

(A)기업은 공정성을 정확성과 특혜 배제라는 원칙을 중심으로 강조하고 알고리즘의 중립성과 데이터 기반 의사결정을 통해 이를 실현하려고 노력하고 있다. 반면, (C)기업은 콘텐츠 추천에서 비즈니스적 전략을 우선시하며 마케팅과 사용자 선호를 고려한 콘텐츠 배치를 통해 이용자 확보에 중점을 두고 있다. (A)기업은 공정성을 기술적 고도화와 내부 검토를 통해 구현하려는 경향이 강한 반면, (C)기업은 사용자 선호와 데이터에 기반하는 차별화된 접근 방식을 보여준다.

3) 책무성

책무성 준수를 위한 (A)기업의 시정 노력은 이루어지고 있지만, 이를 이용자에게 명확히 전달하지 못해 투명성과 공정성 향상을 위한 구체적인 노력을 공지하는 측면에서는 여전히 부족한 것으로 보인다. 인터뷰 결과, (A)기업은 이용자 의견을 수용하기 위한 방법으로 고객센터 문의를 통해 문제를 해결할 수 있는 창구를 운영하고 있지만, 알고리즘 추천 서비스에 특화된 전용 통로는 아직 마련되지 않은 상태임을 확인하였다. 이는 이용자들이 추천 서비스와 관련한 구체적인 질문이나 불만을 보다 쉽게 제기할 수 있는 환경이 미흡하다는 것을 보여준다. 또한, 이용자들이 불만을 제기하기 이전에 선제적으로 질문하거나 우려 사항을 전달할 수 있는 경로에 대해서도 내부적으로 검토가 진행 중이라는 점을 밝혔다. 이는 이용자와의 소통을 강화하고, 책무성을 보다 적극적으로 이행하기 위한 방안을

마련하고자 하는 (A)기업의 의지를 보여주지만, 아직 구체적으로 실행된 사례는 부족한 상황이다. 전반적으로 (A)기업은 이용자 보호와 문제 해결에 대한 체계를 갖추고 있지만, 이를 보다 명확하고 적극적으로 전달하기 위한 전략이 필요하며, 이용자와의 소통 채널 다각화와 선제적 접근이 향후 개선 과제로 남아 있다.

4) 이용자의 선택권 보장

(A)기업은 이용자의 선택권 보장 측면에서 다른 기본원칙에 비해 상대적으로 부족한 노력을 보이고 있는 것으로 나타났으며, 이는 구체적인 운영 현황과 함께 더욱 명확히 드러난다. 현재 운영 방식에서는 이용자의 시청 기록을 기반으로 추천 서비스를 제공하며, 가입 시 초기 선택 사항을 최소화하는 접근을 채택하고 있다. 예를 들어, 스포츠(KBO)와 같은 특정 관심사 내에서 마이팀 설정 기능을 통해 일부 선택권을 제공하고 있으며, 이는 프로필별로 최대 4개까지 설정 가능하도록 운영되고 있다. (A)기업은 이러한 방식이 글로벌 플랫폼인 넷플릭스와는 다른 라이브 콘텐츠, 스포츠, 정기 프로그램 등 제공 콘텐츠의 특성에 맞춘 선택권 제공 방식이라는 점을 강조했다. 이 기업의 현재 접근은 초기 단계에서 장르 선택을 제한하지 않아 콘텐츠 접근성을 높이려는 의도가 반영되어 있다. 그러나 이용자가 콘텐츠 추천 방식에 적극적으로 개입할 수 있는 선택·변경 기능은 부족한 상황으로 평가된다. 이는 이용자가 자신의 콘텐츠 환경을 세밀하게 조정할 수 있는 능동적 선택권 보장에 한계가 있음을 시사한다.

(C)기업은 등급 조정의 방식으로 이용자의 선택권을 보장하고 있다. 계정 설정에서 시청 등급 조정 기능을 제공하여 이용자가 콘텐츠 노출 범위를 통제할 수 있도록 하고 있으며, 기본적으로 계정 생성 시 15세 등급으로 설정된다. 더불어, 프로필별 비밀번호 설정을 통해 개별 사용자 보호를 강화하고 있으며, 특히 아동 및 청소년 보호에 초점을 맞추고 있다. 콘텐츠 추천의 경우 부가적인 서비스로 간주하여 피해를 주지 않는다고 판단하였으며, 가족 내 계정 공유가 자율적으로 이루어지는 만큼 프라이버시와 이용자 보호에 별다른 문제가 없다는 입장을 밝혔다.

(A)기업은 콘텐츠 접근성을 높이는 것을 우선시하면서도 이용자가 추천 서비스에 개입할 수 있는 적극적인 선택권 보장에는 한계를 보이고 있다. 반면, (C)기업은 시청 등급 조정과 같은 기능을 통해 이용자의 콘텐츠 노출 범위를 통제하는 방식으로 선택권을 보장하

며, 특히 아동 및 청소년 보호에 중점을 두고 있다. (A)기업은 제공하고 있는 콘텐츠의 특성에 맞춘 선택권 제공 방식을 모색하고 있는 반면, (C)기업은 계정 설정을 통해 사용자의 프라이버시와 보호를 강조하는 접근을 취하고 있다는 점에서 차이를 보인다.

5) 자율검증 실행

(C)기업은 자율검증 실행을 위해 다양한 국가의 규제 환경을 고려하며, 콘텐츠가 유해하지 않도록 모니터링에 중점을 두고 체계를 운영하고 있다. 콘텐츠의 불필요한 노출을 방지하기 위해 관련 부서에서 지속적으로 모니터링을 수행하고 있으며, 이를 통해 콘텐츠 추천 서비스가 이용자에게 미칠 수 있는 잠재적 위험을 사전에 관리하려는 노력을 기울이고 있다. 이는 글로벌 시장에서 이용자를 보호하고 규제 요구 사항을 준수하는 (C)기업의 전략을 반영한다.

6) 불만 처리 및 분쟁 해결

(A)기업은 불만 처리 및 분쟁 해결과 관련하여, 추천 서비스와 관련된 주요 불만이 시청 기록 삭제 요청에 국한되어 있음을 밝혔다. 현재 시스템에서는 시청 이력을 삭제하더라도 해당 데이터가 알고리즘에 반영된 상태를 유지하고 있으며, 특정 콘텐츠를 비노출 처리하는 기능은 제공되지 않고 있다. 이러한 문제를 개선하기 위해 비노출 기능과 같은 추가 옵션의 제공이 검토되고 있다. 또한, (A)기업은 연령 등급별 콘텐츠 노출 설정 기능을 개발 중이며, 15세 이하 및 유아 사용자 보호를 위한 기능 강화 방안도 논의 중이라고 전했다. 다만, 이러한 보호 기능을 구현하는 과정에서 필요한 정보 수집과 개인정보 보호 간의 상충 문제를 해결하기 위한 윤리적 및 기술적 과제가 남아 있는 상황이다. 고객 통제권 강화를 위해 추천 알고리즘 적용 여부를 이용자가 선택할 수 있는 기능도 현재 검토 중에 있다.

(C)기업은 추천 서비스 품질에 대한 불만이 발생할 가능성을 인정하면서도, 그 외의 불만은 크게 예상되지 않는다는 입장을 보였다. (C)기업은 단순한 1차원적 추천 방식과 등급 조정 기능을 통해 국가별 시청자의 취향 차이를 조정하고 있다고 설명했다. 콘텐츠 자체, 결제, 기능(예: 2배속 재생)과 관련된 불만은 일부 존재하지만, 이용자가 직접 콘텐츠를 선택하는 구조로 인해 추천 서비스와 관련된 불만은 거의 발생하지 않았다고 강조하였다.

(A)기업은 추천 서비스와 관련된 추가 기능과 옵션을 검토 중이며, 연령별 보호 기능 강화에도 초점을 맞추고 있다. 반면, (C)기업은 추천 알고리즘이 간단한 구조를 갖추고 있고

이용자가 직접 콘텐츠를 선택할 수 있는 시스템으로 인해 추천 서비스 관련 불만이 거의 발생하지 않는 상황이다. (A)기업은 시스템 개선과 개인정보 보호 간의 균형을 고민하는 반면, (C)기업은 이용자 자율성을 중심으로 운영되는 구조적 특성에 의존해 불만을 최소화하고 있다.

7) 내부 규칙 제정

(A)기업은 내부 규칙의 측면에서, 내부 회의를 통해 이슈를 해결하고 지침을 마련하는 체계를 유지하고 있다. 이러한 체계는 내부적으로 규칙이 체계적으로 관리되고 있음을 보여주지만, 운영 및 관리에 대한 규칙을 명확히 문서화한 정책 문서는 아직 미비한 상태이다. 또한, 이용자에게 내부 규칙을 구체적으로 공개하는 데 한계가 있어, 내부적으로는 규칙이 체계적으로 관리되고 있음에도 불구하고 외부 이용자와의 소통과 정보 전달 측면에서는 개선이 필요한 것으로 보인다. 이러한 점에서 (A)기업은 내부 규칙의 문서화를 강화하고, 이용자들에게 관련 정보를 보다 명확히 전달하기 위한 개선 노력이 요구된다.

8) 정부 역할

기본원칙 이행을 위한 전반적인 정부의 역할과 관련하여 (C)기업은 투명성과 공정성 제고를 위해 정부가 명확한 가이드라인을 제공함으로써 사업자들에게 책임감을 부여할 필요가 있다는 입장을 밝혔다. 특히, ‘확장성’의 측면에서, 중개 없이 다른 서비스나 광고와 직접 연결하거나 연결 가능성이 있는 사업자는 서비스 증진 외의 목적으로 활용될 우려가 있어 책무와 투명성에 대한 요구가 높아질 수 있음을 강조하며, 이러한 이유로 사업자 특성에 따른 구별이 필요하다고 지적하였다. 또한, 내부 동기가 부족한 사업자들이 책임을 다하지 않을 가능성을 고려해, 입법 강제를 통해 명확한 방향성을 제시하는 것이 효과적일 것이라는 의견을 제시하였다. 나아가, 정부가 제시하는 가이드라인은 신규 사업자들에게 모범 사례로 작용할 수 있으며, 사회적 비판이나 문제에 대응하기 위한 내부 원칙을 설정하는 것이 사업자의 생존에 필수적임을 강조했다. 다만, 이러한 가이드라인은 지나치게 엄격한 수준보다는 현재의 기업들이 충분히 준수할 수 있는 ‘가벼운’ 수준으로 설정하는 것이 적절하다는 의견도 함께 제기되었다.

반면, (A)기업은 정부의 개입이 OTT 사업의 자율성과 경쟁력을 저해할 수 있다는 우려를 표명하였다. 특히, 글로벌 OTT와의 형평성 부족 문제를 지적하며 과도한 규제가 국내

기업에 역차별을 초래할 수 있다는 점을 강조하였다. 필요 지원과 관련하여 글로벌 진출과 번역본 및 더빙본 제공, 또는 기술 개발 지원 확대를 요청하였으며 추천 알고리즘 고도화 연구를 위한 정부 과제와 예산 지원이 필요하다는 입장을 밝혔다. (A)기업은 국내 사업자에 대한 규제가 해외 사업자와 다르게 적용되어 서비스 발전이 저해된다는 점을 지적하며 정부 규제를 최소화하고 해외 OTT와 동등한 수준의 자율성을 부여하는 것이 필요하다고 주장하면서 해외 업체와 국내 업체를 차별하지 않는 정책 방향이 중요하다는 점을 강조하였다. 정부의 개입에 대해 부정적인 태도를 보인 설문 결과와 이를 종합적으로 해석한 바에 따르면, (A)기업은 자율성을 유지하며 사업을 운영하고자 하는 의도가 뚜렷하게 나타난다. 정부의 개입이 OTT 사업자의 자율성을 침해하거나 경쟁력을 저하시킬 수 있다는 우려를 표명한 (A)기업은, 기업 스스로 책임을 다할 수 있는 환경 조성의 중요성을 강조하고 있다.

(C)기업은 정부의 명확한 가이드라인이 투명성과 공정성을 강화하고 신규 사업자들에게 모범 사례로 작용할 수 있는 중요한 도구라고 평가하며 규제와 지원의 균형을 강조했다. 반면, (A)기업은 정부 개입이 자율성을 침해하고 국내 OTT 기업의 경쟁력을 약화시킬 수 있다고 보며, 해외 사업자와의 공정한 경쟁 환경 조성을 우선시하고 있다. (C)기업은 정부의 개입을 긍정적 도구로 보며 이를 통해 사업자 책임감을 강화해야 한다는 입장이 뚜렷한 반면, (A)기업은 자율성을 유지하며 스스로 책임을 다할 수 있는 여건을 마련하는 것이 중요하다고 강조하며 두 기업 간 접근 방식에 뚜렷한 차이를 보였다.

제 5 절 기대효과

본 연구는 인공지능 기반 미디어 추천 서비스의 이용자 보호를 위해 마련된 기본원칙의 이행 수준을 파악하고, 이를 토대로 정책 개선 방향을 제시하는 것을 목표로 하였다. 본 연구의 기대효과는 크게 네 가지로 요약할 수 있다.

1. 미디어 플랫폼의 이용자 보호 원칙 준수 여부 평가

본 연구는 주요 미디어 플랫폼의 인공지능 기반 추천 서비스가 투명성, 공정성, 책무성

의 핵심 원칙과 정보 공개, 선택권 보장, 자율검증 실행, 불만 처리 및 분쟁 해결, 내부 규칙 제정의 실행 원칙을 얼마나 효과적으로 이행하고 있는지를 체계적으로 검증함으로써, 각 플랫폼의 운영 현황을 명확히 파악할 수 있었다. 이를 위해 설문조사와 인터뷰, 서비스 약관 및 고객센터 웹페이지 분석을 통해 사업자의 실제 운영 방식을 다각적으로 분석하였다. 이로써 플랫폼별 이행 수준의 차이와 개선이 필요한 영역을 진단하고, 주요 플랫폼의 우수 사례와 모범 운영 방식을 발굴할 수 있었다. 특히, 서비스 약관과 고객센터 웹페이지에 명시된 이용자 불만 처리 절차와 추천 알고리즘의 투명성 확보 수준을 점검함으로써, 사업자가 외부에 공시한 정책과 실제 서비스 운영 간의 일치 여부를 평가할 수 있었다. 이러한 분석 결과는 향후 정책 수립에 중요한 기초자료로 활용될 수 있으며, 사업자의 자율적 개선을 유도하는 데에도 기여할 것으로 기대된다.

2. 설문조사 및 심층인터뷰 기반 정책적 시사점 도출

본 연구는 정량적·정성적 데이터를 기반으로 미디어 플랫폼의 이행 수준을 평가하여 정책적 개선 방향을 제시한다. 구체적으로는, 설문조사와 심층 인터뷰에서 도출된 사업자의 의견을 종합하여 플랫폼의 자율적 이행을 독려하기 위한 가이드라인과 법적 규제 방안을 제시함으로써, 정부와 규제기관이 투명성, 공정성, 책무성을 강화하기 위한 실질적 정책 수단을 마련하는 데 기여할 것으로 기대된다. 특히, 본 연구에서 제시하는 정책적 시사점은 법적 규제뿐만 아니라, 자율적 개선을 유도하기 위한 정부의 지원 정책과 가이드라인 수립에도 활용될 수 있다. 자율검증 체계의 운영 방식, 불만 처리 및 분쟁 해결 절차의 개선 방안 등은 사업자가 직접 개선할 수 있는 영역으로, 정책적 시사점으로 제시되었다. 이는 향후 관련 법령 개정이나 가이드라인 제정 시 중요한 참조 자료가 될 수 있을 것이다.

3. 이용자 권의 보호와 미디어 다양성 확보에 기여

본 연구는 미디어 플랫폼의 추천 알고리즘이 투명성, 공정성, 책무성을 확보할 수 있도록 개선됨에 따라, 이용자의 권익 보호와 미디어의 다양성 확보에 기여할 수 있을 것으로 기대된다. 알고리즘의 투명성이 강화되면 이용자는 자신의 데이터가 어떻게 활용되는지 더 잘 이해하고, 이를 통제할 수 있는 권리를 보장받을 수 있다. 또한, 플랫폼이 공정성을

강화하여 알고리즘 편향성을 최소화함으로써, 다양한 콘텐츠에 접근할 기회가 확대될 것이다. 본 연구에서는 이용자 선택권 보장 원칙의 이행 실태를 분석하여, 사용자가 추천 알고리즘의 매개변수를 직접 조정할 수 있는 기능(예: 시청 이력의 비중 조절 등)이 부족함을 지적하였다. 이러한 분석 결과는 플랫폼의 알고리즘 편향을 줄이고 다양한 콘텐츠의 노출을 보장할 수 있는 정책 방안의 기초자료로 활용될 수 있다. 이용자 선택권 보장, 자율검증 체계의 강화, 미성년자 보호 조치의 필요성 등은 모두 미디어의 다양성 확보와 이용자 권익 보호에 기여할 수 있는 중요한 요소로 작용할 것으로 기대된다.

4. 미디어 플랫폼의 신뢰성 제고에 기여

본 연구에서 제안하는 투명성, 공정성, 책무성의 핵심 원칙과 실행 원칙이 체계적으로 이행될 경우, 미디어 플랫폼의 서비스 신뢰성이 향상될 수 있다. 알고리즘의 투명성이 강화됨으로써 사용자 신뢰가 높아지고, 공정성을 확보함으로써 추천 알고리즘의 편향성이 감소하며, 책무성이 강화됨으로써 불만 처리 및 분쟁 해결의 접근성이 높아질 것이다. 예를 들어, 본 연구에서는 사업자들이 추천 알고리즘의 운영 방식을 이용자에게 명확히 공개하는 방법을 분석하였고, 이를 통해 알고리즘 설명 방식 개선의 필요성을 제안하였다. 특히, 서비스 이용 약관과 고객센터의 불만 처리 메커니즘을 검토하여, 불만 접수 및 처리 절차에 대한 구체적인 개선 방안을 제시하였다. 이러한 변화는 사용자 경험의 질을 향상시키고, 플랫폼의 장기적인 성장과 지속 가능한 운영을 가능하게 하는 요소로 작용할 것이다. 특히, 신뢰성 제고는 플랫폼 사업자가 글로벌 경쟁에서 우위를 확보할 수 있는 중요한 전략적 이점으로 작용할 수 있다.

제5장 메타버스 및 알고리즘 사업자 규제 이행 실태 연구

제1절 연구 배경 및 목적

1. 연구 배경

디지털 기술의 급격한 발전과 함께 메타버스가 빠르게 확산되고 있으며, 다양한 산업에서 필수 요소로 활용되고 있다. 메타버스는 단순한 가상현실을 넘어 교육, 엔터테인먼트, 비즈니스 등 다양한 분야에서 활용되면서 혁신적인 가치를 창출하고 있다(김정수, 2023; 신현주, 2022). 교육 분야에서는 가상 교실을 통해 지리적 제약 없이 학생들에게 몰입감 있는 학습 경험을 제공하며(조연직 외, 2022), 의료, 항공, 군사 등 전문 분야에서는 시뮬레이션 기반 훈련을 통해 비용 절감과 안전성 향상을 실현하고 있다(이하은, & 한정엽, 2021). 엔터테인먼트 산업에서는 가상 콘서트와 이벤트를 통해 전 세계 팬들과 소통하며 새로운 수익 모델을 창출하고 있으며, 게임과 소셜 플랫폼을 통해 사용자 간 상호작용을 극대화하고 있다(이연경 외, 2022; 정지은 외, 2022).

비즈니스 영역에서도 메타버스의 활용이 두드러지고 있다. 기업들은 가상 회의실을 통해 효율적인 원격 협업을 실현하고 있으며, 부동산 및 건축 분야에서는 가상 부동산 거래와 건축 시뮬레이션을 통해 새로운 비즈니스 모델을 창출하고 있다(신현주, 2022). 또한 메타버스를 새로운 마케팅과 판매 채널로 활용하여 소비자와의 접점을 확대하고, 독창적인 소비자 경험을 제공하고 있다. 특히, 메타버스를 통해 실시간 협업, 가상 상품 판매, 몰입형 학습 등 다양한 서비스가 가능해지고 있어, 미래 경제의 핵심 플랫폼으로 주목받고 있다.

국내의 주요 기업들도 메타버스 시장의 잠재력을 인식하고 적극적으로 참여하고 있다. 엔비디아의 옴니버스 플랫폼, 로블록스의 사용자 생성 콘텐츠 기반 경제 시스템, 메타의 호라이즌 월드, 마이크로소프트의 메시와 팀즈 등은 각 기업의 강점을 살려 메타버스 생태계를 구축하고 있다. 국내에서도 SK텔레콤의 이프랜드, 직방의 메타폴리스 등 자체 메타버스 플랫폼 개발이 활발히 이루어지고 있으며, 다양한 기업들이 제페토와 같은 기존

플랫폼을 활용해 새로운 마케팅과 비즈니스 기회를 모색하고 있다.

그러나 메타버스 서비스의 확산은 동시에 우리 사회가 해결해야 할 새로운 과제들을 제시하고 있다. 개인정보 보호와 데이터 안전성 문제가 주요한 사회적 이슈로 떠오르고 있으며, 특히 잘못된 알고리즘 설계는 편향, 차별, 프라이버시 침해 등의 심각한 부작용을 일으킨다.

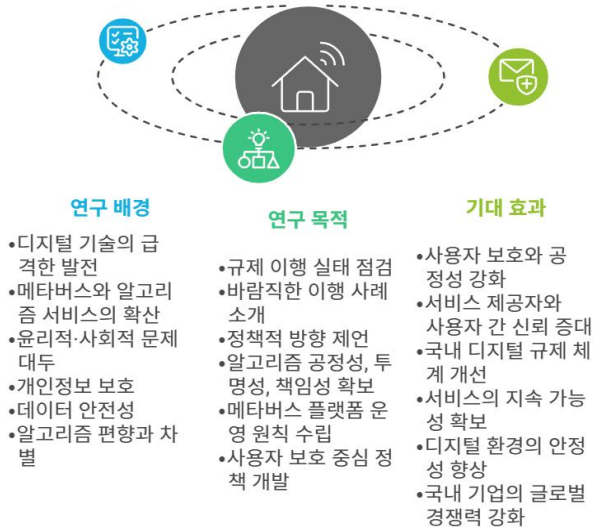
메타버스를 사용하는 플랫폼에서는 사용자의 세부적인 활동 데이터가 수집되고 활용되는데, 이는 프라이버시 침해의 위험을 높이며, 데이터의 오남용 가능성을 증가시킨다(김정환, 2023; 유승현, 2022). 특히 복합적인 통합 데이터베이스는 사생활 보호나 정보 보안을 위한 안전장치가 거의 전무한 채로 개인의 가장 민감한 정보를 수집하고 있어 많은 우려를 낳고 있기도 하다.

이러한 문제들은 메타버스의 지속 가능성을 위협하는 요인으로 작용한다. 사용자들의 신뢰를 잃게 되면 서비스 이용률이 감소하고, 결국 플랫폼의 존속 자체가 위협해질 수 있다. 따라서 사용자 보호와 사회적 책임을 동시에 고려하는 규제와 정책적 지원이 필요하다. 단순히 기술적 혁신에만 초점을 맞추는 것이 아니라, 서비스의 사회적 영향을 고려하여 공정성과 투명성을 보장하는 방향으로 나아가야 한다. 메타버스는 기술적 도구를 넘어 사회적 영향력이 큰 플랫폼으로 자리 잡고 있기 때문에, 사회적 책임을 다하는 운영 원칙이 더욱 중요해지고 있다(이희욱, 2023).

미국과 유럽 등 주요 국가들은 이러한 문제를 해결하기 위해 개인정보 보호법, 디지털 서비스법 등 다양한 규제를 강화하고 있다. 예를 들어, 유럽연합의 GDPR(일반 개인정보 보호법)은 데이터 처리와 보호에 대한 강력한 기준을 제시하고 있으며, 미국은 플랫폼 기업의 책임성을 높이기 위해 디지털 서비스에 대한 규제를 강화하고 있다. 국내에서도 이러한 국제적 흐름에 발맞추어 관련 법과 규제를 마련하고, 국제적 기준에 부합하는 체계적 대응이 요구되고 있다.

〔그림 5-1〕 연구의 개요

연구의 배경, 목적, 기대효과



2023년 11월 29일에는 『메타버스 이용자 보호 기본원칙』을 발표했다(방송통신위원회, 2023). 이 기본원칙은 메타버스 이용자들이 안전하고 편리하게 서비스를 이용할 수 있도록 마련된 자율규범으로, 메타버스 플랫폼 서비스 제공 사업자에게 권고되는 내용을 담고 있다. 주요 내용으로는 6가지 기본원칙이 제시되었는데, 이는 공동체 가치 형성과 보호, 이용자의 서비스 이해와 의견 제시 보장, 이용자 간 갈등 조정, 공정한 경제활동 보호, 데이터 통제권 보장, 지속가능발전 지향 등을 포함한다. 특히 메타버스에서 유통되는 디지털 재화와 서비스의 공정한 거래, 지식재산권 보호, 이용자 데이터 관리 권리 보장 등이 강조되었다. 또한 아바타에 대한 성추행과 스토킹 제한, 사이버 불링 신고 및 제재, NFT 구매자 이전 권리 등 구체적인 이용자 보호 방안을 담은 실천 규범 안도 함께 제시했다.

2. 연구 목적

본 연구는 메타버스의 규제 이행 실태를 종합적으로 점검하여 이용자 보호 기본원칙의 실효성을 평가하고 정책적 시사점을 도출하고자 한다. 구체적으로, 본 연구는 다음과 같은 세부 내용을 수행할 것이다. 첫째, 메타버스 플랫폼 서비스 제공자들의 이용자 보호 기본원칙 준수 여부를 체계적으로 조사하고 분석한다. 이를 통해 현재의 자율규제 체계의 효과성을 평가하고, 개선이 필요한 영역을 식별하고자 한다. 둘째, 이용자 보호 기본원칙의 실제 적용 과정에서 발생하는 문제점과 한계를 파악하여, 보다 실효성 있는 규제 방안을 모색한다. 마지막으로, 연구 결과를 바탕으로 이용자 보호를 강화하고 디지털 환경의 신뢰성을 제고하기 위한 구체적인 정책 제언을 제시할 것이다.

궁극적으로 본 연구는 이용자의 권익을 보호하고, 공정하고 투명한 디지털 서비스 환경을 조성하며, 나아가 지속 가능한 디지털 생태계 구축에 기여하는 것을 목표로 한다. 이를 통해 메타버스가 제공하는 혁신적 가치를 극대화하면서도, 발생 가능한 부작용을 최소화하는 균형 잡힌 접근을 모색하고자 한다.

3. 기대효과

본 연구는 메타버스의 규제 이행 실태를 분석함으로써 사용자 보호와 공정성 강화에 실질적인 기여를 할 것으로 기대된다. 연구를 통해 도출된 결과는 데이터 보호와 윤리적 운영 기준을 강화하여 사용자와 서비스 제공자 간 신뢰를 증대시키는 데 도움을 줄 것이다. 이는 궁극적으로 서비스의 지속 가능성을 높이고, 디지털 환경에서의 안정성을 확보하는데 기여할 것이다.

알고리즘의 공정성과 투명성을 확보하는 방안을 제시함으로써, 서비스 제공자가 사용자 맞춤형 추천 시스템을 보다 책임감 있게 운영하는 데 기여할 것이다. 이는 편향적 결정과 차별적 결과를 방지하고, 사용자가 공정한 디지털 환경에서 서비스를 이용할 수 있는 기반을 마련하고 이를 통해 사용자 경험의 질을 높이고, 디지털 서비스의 사회적 책임을 강화할 수 있을 것이다. 이는 궁극적으로 플랫폼의 신뢰성과 안정성을 강화할 수 있을 것이다. 공정한 이용 환경이 보장되면, 사용자 간의 갈등이 줄어들고, 메타버스가 사회적 관계 형성과 경제 활동의 중심 플랫폼으로 자리 잡는 데 긍정적인 영향을 미칠 것이다.

또한 단순히 규제와 정책적 시사점에 그치지 않고, 사용자 경험을 개선하고, 디지털 기술이 긍정적인 사회적 영향을 미치도록 유도하는 데 활용될 수 있다. 특히, 사용자 신뢰를 증대시키는 디지털 서비스의 발전은 서비스 제공자와 사용자 모두에게 장기적인 혜택을 제공할 것이다.

마지막으로, 메타버스의 사회적 역할과 책임을 강조함으로써, 디지털 기술이 공동체 가치를 형성하고 유지하는 데 기여할 수 있는 방향성을 제시할 것이다. 이는 기술의 활용이 단순한 효율성을 넘어, 인간 중심의 기술 환경을 조성하는 데 도움을 줄 것이다. 또한 디지털 생태계의 지속 가능성을 확보하고, 기술적 진보와 사회적 책임이 조화를 이루는 사례를 제시함으로써, 미래의 디지털 환경 설계에 대한 청사진을 제공할 것이다. 이러한 연구 결과는 학계와 산업계, 정책 입안자들에게 중요한 참고자료가 될 것으로 기대된다.

제 2 절 메타버스 서비스 활용 사례

1. 메타버스에 대한 정부의 제도적 지원

가. 국내의 제도적 지원

정부도 메타버스 산업의 활성화를 위해 다양한 방안을 추진하고 있다. 지난 2022년 1월, 과학기술정보통신부는 ‘디지털 신대륙, 메타버스로 도약하는 대한민국’을 비전으로 설정한 메타버스 신산업 선도전략 발표하고 2026년까지 글로벌 메타버스 시장 점유율 5위 달성을 목표로 제시했다(과학기술정보통신부, 2022). 이를 위해 메타버스 전문가 4만 명 양성, 매출액 50억 원 이상 전문기업 220개 육성, 메타버스 모범사례 50건 발굴 등과 같은 세부목표를 제시하기도 했다. 2023년 3월에는 메타버스 산업 진흥법 제정 추진, 자율규제 및 임시기준 제도 도입, 메타버스 윤리원칙 확산 및 교육 확대, 지식재산권 제도 정비로 메타버스 서비스 활성화, 이동형 영상정보처리기를 통한 개인영상정보의 안전한 활용기준 마련 등의 방안이 담긴 메타버스 생태계 활성화를 위한 ‘가상융합세계(메타버스) 생태계 활성화를 위한 선제적 규제혁신 방안’을 발표하기도 했다(과학기술정보통신부, 2023).

2024년 2월에는 정보통신산업진흥원이 주최로 ‘2024년 메타버스산업 통합 사업설명회’를 개최하고 메타버스 산업의 글로벌 시장 선도를 위해 플랫폼 개발지원, 인력양성, 기업

육성, 기술개발, 저변확산 등 다양한 맞춤형 지원 사업을 추진하기로 발표했다(정보통신산업진흥원, 2024). 구체적으로, 청년 인재 양성을 위해 메타버스 아카데미와 융합대학원 운영에 약 187억 원을 투입하고, 지역특화산업과 연계한 콘텐츠 실증과 사업화를 위해 메타버스 허브센터를 구축하기로 했으며, 홀로그램 기술 개발과 사업화 실증 및 메타버스 콘텐츠와 디바이스 개발을 위한 다양한 지원센터를 운영을 추진하기로 했다. 이와 함께 수출 가능성이 높은 국내 중소기업의 해외 진출을 지원하기 위해 맞춤형 컨설팅과 글로벌 행사 참여를 제공하고, 500억 원 규모의 디지털콘텐츠코리아펀드를 조성해 유망 기업에 투자하기로 했다. 또한 XR 기술과 인터페이스 기술, 홀로그램 기술 개발에 약 225억 원을 투입해 기술 경쟁력을 확보하고 가상융합산업진흥법 시행에 따라 하위 법령을 마련하여 메타버스 윤리원칙을 확산해 안전한 환경을 조성할 계획을 발표했다.

2024년 2월에는 메타버스 산업을 지원하고 관련 규제를 개선하기 위한 ‘가상융합산업진흥법’ 제정안이 국무회의에서 의결되었다. 가상융합산업 진흥법 시행령안은 메타버스 산업을 육성하기 위해 마련된 법령으로, 체계적인 진흥 정책 추진, 자율규제 환경 조성, 선제적 규제개선을 위한 임시기준 제도 마련 등을 주요 내용으로 한다. 시행령에 따라 가상융합산업 진흥을 위한 기본계획 수립, 산업 현황 및 규제개선 과제 발굴을 위한 실태조사 실시, 전문인력 양성, 지역별 가상융합산업지원센터 지정 등이 이루어지고 있다. 또한, 가상융합사업자들은 정보 제공, 상담, 사업 공간 제공, 유통 활성화 지원 등 다양한 행정적·재정적 지원을 받을 수 있게 되며, 중소 가상융합사업자를 우선 지원할 수 있는 근거도 마련되었다. 이 법령은 가상융합기술·서비스 개발과 관련해 우선허용·사후규제 원칙을 명문화하여 혁신을 저해하지 않도록 하는 동시에, 신산업 영역에서 발생할 수 있는 이용자 보호 이슈에 대응하기 위한 민간 중심의 자율규제 근거를 마련했다. 또한, ‘가상융합산업 영향평가’ 제도를 도입하여 중앙행정기관과 지자체가 가상융합사업을 추진할 때 민간 시장에 미치는 영향을 분석하도록 하여 메타버스 산업의 체계적인 진흥과 규제 혁신, 글로벌 메타버스 시장에서 한국의 경쟁력을 제고할 것으로 기대되고 있다.

나. 해외의 제도적 지원

메타버스 산업의 중요성이 부각되면서, 여러 주요 국가들도 메타버스 산업 활성화를 위해 다양한 정책을 추진하고 있다. 미국은 메타버스의 핵심 기술로 여겨지는 XR(확장 현실)을 국가 전략 기술로 지정하고 이를 포함한 10대 핵심 기술 영역에 대한 연구개발(R&D)을

지원하고 있다. 특히, 2022년 발표된 ‘반도체와 과학법’에서는 XR 관련 기술의 개발과 혁신을 전략적으로 육성하기 위해 국립표준기술연구소(NIST)와 국립과학재단(NSF) 등을 통해 대규모 투자를 진행하고 있다. 미국 증권거래위원회(SEC)는 메타버스 내 가상 자산에 대해 증권법을 적용할 수 있다는 입장을 밝히기도 했다(한상열, 2024).

중국은 2023년 9월 공업정보화부를 포함한 4개 부처가 공동으로 ‘메타버스 산업의 혁신 발전을 위한 3개년 행동 계획(2023-2025)’을 발표했다(남현숙 & 곽나연, 2023). 이 계획은 2025년까지 메타버스 기술, 산업, 앱(APP) 등 다양한 영역에서 획기적인 발전을 이루어 디지털 경제 성장의 중요한 엔진 역할을 하도록 만들겠다는 목표를 설정하고 있으며, 글로벌 영향력을 갖춘 메타버스 관련 기업 3~5개를 육성하고, 산업 클러스터 단지 3~5개를 조성하겠다는 계획을 포함하고 있다. 또한 핵심 메타버스 기술의 통합적인 혁신을 강화하기 위한 노력이 필요하다고 강조하며, 산업 메타버스와 디지털 생활 등 다섯 가지 주요 과제를 제시하고 있다. 메타버스 산업의 성장을 촉진하기 위해 첨단기술 확보와 산업 시스템 구축 등 세부적인 조치를 강조하고 있으며, 이는 중국 정부가 메타버스를 국가 차원에서 적극적으로 지원하고 육성하려는 의지를 보여주는 것이다. 중앙정부의 정책 기조에 맞춰 다양한 지방 정부들도 메타버스 산업 육성 정책을 발표하며 범국가적 차원에서 메타버스의 발전을 촉진하고 있다.

유럽연합은 메타버스와 가상 세계에 대한 포괄적인 접근 방식을 취하고 있으며, 2023년 7월 유럽 위원회가 발표한 ‘Web 4.0 및 가상 세계에 대한 새로운 전략’이 그 중심에 있다(TaylorWessing, 2023). 이 전략은 EU 시민들의 삶과 일, 상호작용 방식을 변화시킬 잠재력을 가진 메타버스를 EU의 가치와 기본권을 반영하도록 하는 것을 목표로 한다. EU는 Web 4.0 및 가상 세계 전문가 수를 늘리고 일반 대중의 이해를 높이기 위한 ‘가상 세계 톨박스’를 만들 계획이며, 유럽의 가상 세계 산업 생태계를 지원하기 위해 새로운 유럽 파트너십을 구축하고 규제 샌드박스를 통해 기술과 서비스를 테스트할 수 있는 환경을 조성할 예정이다.

또한 EU와 회원국 간의 긴밀한 협력을 통해 가상 세계가 가져올 사회적 변화에 대응하고, 공정하고 평등한 거버넌스를 위해 개방형 표준을 지원할 계획이다. 유럽 위원회는 현재 메타버스에 대한 새로운 규제를 제안하기보다는 기존의 디지털 기술 관련 법규가 Web 4.0과 가상 세계에도 적용될 수 있다고 보고 있다. 더불어 유럽 의회는 2024년 10월 메타

버스에서의 지적재산권에 대한 결의안을 채택하여, EU의 지적재산권 보호 법규가 가상 세계에도 완전히 적용된다는 점을 명확히 했다. 이러한 정책들을 통해 유럽연합은 메타버스 기술 발전에서 주도적인 역할을 하고, 동시에 EU의 가치와 원칙이 반영된 안전하고 혁신적인 가상 세계를 구축하고자 노력하고 있다.

일본은 메타버스 산업 발전을 위해 정부와 민간 기업이 적극적으로 나서고 있다. 2022년 6월, 일본 정부는 “지적재산 진흥계획 2022”를 통해 메타버스, NFT, 생성형 AI 등 신기술에 대한 대응을 강조하며, 이에 따라 메타버스의 법적 문제를 다루는 가이드라인 개발을 계획하고 있다(남현숙 & 궤나연, 2023). 일본 총무성은 2022년 8월 메타버스 활용에 관한 전담 연구그룹을 출범시켰고, 2023년 7월 포괄적인 보고서를 발표했다. 이 보고서는 메타버스를 “인터넷 또는 다른 네트워크를 통해 접근 가능한 가상 디지털 공간으로, 사용자들이 소통할 수 있게 하는 것”으로 정의하고 있다.

일본 내각부는 2022년 11월 메타버스 콘텐츠 관련 법적 문제를 다루는 민관 협력 위원회를 설립했고, 2023년 5월 종합 보고서를 발표했다. 이 보고서는 메타버스 내 실제 브랜드 상품의 무단 모방과 판매 문제 등을 다루고 있다.

민간 부문에서는 일본 메타버스 경제 구역 구축을 위한 노력이 진행 중인데, 미즈호 금융그룹, 스미토모 미쓰이 금융그룹, 미쓰비시 UFJ 금융그룹 등 주요 기업들이 참여하여 ‘RYUGUKOKU’라는 메타버스 플랫폼을 개발하고 있다.

시장 전망에 따르면, 일본의 메타버스 산업은 2023년 133억 4,320만 달러 규모에 이를 것으로 예상되며, 2023년부터 2030년까지 연평균 36.5%의 성장률을 보일 것으로 전망되고 있다. 일본 정부는 Web3 서비스 확대를 위한 노력을 지원하고 있으며, 세제 혜택 등을 통해 디지털 미래를 수용하는 기업들을 지원할 계획을 밝혔다.

〈표 5-1〉 국가별 주요 정책 및 지원 내용

국가	주요 정책 및 지원 내용
한국	• 2022년: ‘메타버스 신산업 선도전략’ 발표 (2026년까지 글로벌 시장 점유율 5위 목표)
	• 2023년: ‘가상융합세계 생태계 활성화를 위한 선제적 규제혁신 방안’ 발표
	• 2024년: ‘메타버스산업 통합 사업설명회’ 개최 (인력양성, 기업육성, 기술개발 등 지원)
	• 2024년: ‘가상융합산업 진흥법’ 제정안 국무회의 의결
미국	• XR을 국가 전략 기술로 지정
	• ‘반도체와 과학법’을 통해 XR 관련 기술 개발 및 혁신 지원
	• NIST와 NSF를 통한 대규모 투자 진행
중국	• ‘메타버스 산업 혁신 발전을 위한 3개년 행동계획(2023-2025)’ 발표
	• 지방정부 차원의 스타트업 지원, 인프라 확충, 규제 개선
유럽연합	• 가상세계 혁신 전략 제시
	• 메타버스 글로벌 표준화 선도 노력
	• 디지털 규범 정립 및 R&D 지원 강화
일본	• 메타버스와 웹 3.0 산업 육성을 위한 과제 발굴
	• 블록체인, NFT 등 연계 기술 개발 지원
	• 기업의 메타버스 관련 혁신적 아이디어 실현 지원

자료: 연구진 작성

2. 메타버스 활용 사례

메타버스의 활용 사례를 교육, 엔터테인먼트, 비즈니스 및 쇼핑, 헬스케어, 공공서비스 각 분야로 구분하여 살펴보았다.

가. 교육

교육 분야에서 메타버스는 가상현실(VR)과 증강현실(AR)을 통해 실제와 유사한 학습 환경을 제공하여 학습자가 몰입감 있게 경험할 수 있도록 돕는다. 이를 통해 복잡한 개념이나 추상적인 주제를 시각적으로 쉽게 이해할 수 있으며, 학습자는 아바타를 활용해 자신의 속도와 스타일에 맞춘 맞춤형 학습을 진행할 수 있다. AI와 데이터 분석 기술을 활용하면 학습자의 필요에 따라 콘텐츠를 조정하는 것도 가능하다. 또한, 가상 환경은 실제 실험

이나 현장 학습보다 비용이 적게 들고 위험 요소를 줄여주어 안전성을 확보할 수 있으며, 위험한 화학 실험이나 공학 설계를 시뮬레이션하는 데 유용하다. 나아가 메타버스는 지리적 한계를 초월해 전 세계 학습자들이 실시간으로 협업하고 소통할 수 있는 글로벌 교육 환경을 조성하는 데 기여한다(조연직 외., 2022).

국내 어떤 대학교는 메타버스 플랫폼을 통해 캠퍼스 투어와 학과 소개를 진행하며, 신입생들이 가상공간에서 학교를 체험할 수 있도록 지원하고 있다. 이와 비슷하게 최근 SK 텔레콤은 ‘이프랜드(ifland)’ 플랫폼을 통해 한 대학교의 입학식을 메타버스에서 진행했다. 초중등 교육에서도 메타버스를 활용하여 가상공간에서 유적지를 탐방하거나 역사적 사건을 체험하는 등 학습에 적극 활용하고 있다. 과학기술정보통신부도 디지털 인재 양성 프로그램의 일환으로 메타버스를 활용한 코딩과 소프트웨어 교육을 제공하며, IT 기술에 대한 몰입형 학습 기회를 확대하고 있다.

사설 교육에서도 메타버스 활용이 늘어나고 있는데, 한 교육 스타트업은 가상 실험실을 구축해 화학과 물리학 실험을 안전하게 진행할 수 있는 학습 환경을 제공하고 있으며, 이로 인해 실험 비용 절감과 안전성이 확보되고 있다. 또 다른 사례로, 언어 학습 플랫폼에서는 학생들이 가상 세계에서 영어를 실제 대화처럼 연습할 수 있는 프로그램을 운영하며 학습 효과를 높이고 있다. 이러한 사례들은 메타버스가 국내 학습 환경에 혁신적 변화를 가져오고 있음을 보여준다.

해외에서는 메타버스를 활용한 학습이 다양한 분야에서 이루어지고 있다. 미국의 zSpace 플랫폼은 학생들이 가상공간에서 해부학 실습이나 물리학 실험을 할 수 있도록 지원하며, STEM(과학, 기술, 공학, 수학) 교육을 혁신적으로 변화시키고 있다(AET Labs, 2024). 영국의 Imperial College London은 메타버스를 활용해 의학 교육을 진행하고 있으며, 학생들이 실제 환자를 다루기 전 수술을 연습할 수 있는 가상 수술실을 제공하고 있다(Coleman, 2023). 중국에서는 언어 학습 플랫폼인 GidiX를 통해 학습자들이 가상 환경에서 영어를 자연스럽게 사용하는 경험을 제공하며, 언어 교육에서 메타버스의 효과를 입증하고 있다(남현숙, 궤나연, 2023).

호주의 한 교육 기관은 가상현실을 활용해 농업 훈련을 제공하며, 학습자들이 농업 기술과 기계를 안전하게 익힐 수 있도록 지원하고 있다. 캐나다의 일부 학교는 메타버스를 활용해 원주민 역사와 문화를 가상으로 체험할 수 있는 프로그램을 운영하며, 학생들이

보다 몰입감 있게 문화적 이해를 높일 수 있는 기회를 제공하고 있다. 이러한 사례들은 메타버스가 교육의 경계를 확장하고 학습 경험을 혁신적으로 변화시키는 도구로 자리 잡고 있음을 보여준다.

이처럼 메타버스는 몰입형 학습 환경, 개인화된 경험, 안전성과 비용 효율성, 글로벌 협업 등 다양한 장점을 통해 교육 분야에서 혁신적인 변화를 이끌고 있으며 이와 관련된 여러 사례들은 메타버스의 실제 활용 현황과 가능성을 보여주며, 미래 교육의 방향성을 제시하고 있다.

나. 엔터테인먼트

엔터테인먼트 분야에서의 활약상도 대단하다. 국내 네이버의 ‘제페토’ 플랫폼은 사용자들이 3D 아바타를 통해 가상공간에서 소통하고, K-pop 아티스트들과의 협업을 통해 글로벌 팬들과의 교류를 촉진하고 있다. SK텔레콤의 ‘이프랜드’는 가상 회의와 소셜 네트워킹을 위한 메타버스 플랫폼으로, 다양한 가상공간에서 만남과 행사를 개최할 수 있도록 지원하고 있다. 하이브의 ‘위버스’는 아티스트와 팬들이 소통할 수 있는 온라인 플랫폼으로, 글로벌 K-pop 아티스트들과의 협업을 통해 팬덤 문화를 확장하고 있으며, SM엔터테인먼트의 ‘디어유 버블’은 아티스트와 팬들이 1:1 채팅을 할 수 있는 서비스를 제공하고 있다. 엔씨소프트의 ‘유니버스’는 다양한 아티스트들의 콘텐츠를 제공하며, 팬들이 가상공간에서 아티스트와 소통할 수 있는 플랫폼을 운영하고 있다.

해외에서도 메타버스를 활용한 엔터테인먼트 사례가 활발히 전개되고 있다. 미국의 ‘로블록스’는 메타버스 기반의 게임 플랫폼으로, 다양한 게임과 가상 콘서트를 통해 사용자에게 새로운 경험을 제공하고 있다. 미디어사 iHeartMedia는 로블록스에 ‘iHeartLand’를 구축하여 가상 콘서트와 팟캐스트를 개최하며, 새로운 형태의 엔터테인먼트를 선보이기도 했다. 닌텐도의 ‘모여봐요 동물의 숲’은 가상 섬에서 생활하며 다른 사용자와 교류할 수 있는 게임으로, 팬데믹 기간 동안 큰 인기를 끌었다. 페이스북은 메타버스 플랫폼 ‘호라이즌 월드’를 통해 사용자들이 가상공간에서 다양한 활동을 할 수 있도록 지원하고 있으며, 에픽게임즈의 ‘포트나이트’는 가상 콘서트와 이벤트를 통해 게임 이상의 경험을 제공하며, 메타버스의 가능성을 보여주고 있다.

특히 2020년 4월 미국의 힙합 아티스트인 Travis Scott의 가상 콘서트는 메타버스를 활

용하여 콘서트를 개최한 유명한 예이다. Fortnite라는 메타버스 플랫폼에서 진행된 이 가상 콘서트에서는 Travis Scott의 디지털 아바타가 등장해 공연을 진행하는 동안 플레이어들은 공연 중 음악에 맞춰 점프하거나 댄스를 이동하며 자유롭게 콘서트를 체험하는 활동이 가능했다. 또한 공연은 최신 게임 기술과 그래픽으로 제작되어 플레이어들이 극대화된 몰입감을 경험하기도 했다. 이 가상 공연은 전 세계 약 2,780만 명의 플레이어들이 참여하여 메타버스와 가상 이벤트의 가능성을 보여주는 상징적인 사례가 되었다.

다. 비즈니스 및 쇼핑

비즈니스와 쇼핑에서도 메타버스를 활용한 다양한 사례가 주목받고 있다. 롯데홈쇼핑은 가상 모델 '루시'를 활용해 메타버스에서 제품을 소개하고, MZ세대를 겨냥한 마케팅을 강화하고 있다. GS리테일은 메타버스 플랫폼 '시나몬'에 가상 GS25 점포를 열어, 아바타를 통해 기프티콘을 구매하고 이를 오프라인 매장에서 사용할 수 있는 서비스를 운영 중이다. 현대자동차는 '제페토' 플랫폼에 가상 전시장을 구축해 신차 모델을 소개하고, 사용자들이 가상 시승을 체험할 수 있도록 했다.

신세계백화점은 메타버스를 활용한 가상 매장을 운영하며, 고객들이 아바타를 통해 제품을 체험하고 구매할 수 있는 환경을 제공하고 있다. LG유플러스는 가상 오피스를 구축해 직원들이 메타버스 공간에서 회의와 협업을 진행할 수 있도록 지원하고 있다. 이처럼 국내 기업들은 메타버스를 통해 고객 경험을 혁신하고, 비즈니스 효율성을 높이는 다양한 시도를 하고 있다.

해외에서도 메타버스가 비즈니스와 쇼핑의 혁신적 도구로 활용되고 있다. 월마트는 '로블록스' 플랫폼에 가상 매장을 개설해, 고객들이 가상 현실에서 제품을 체험하고 구매할 수 있는 경험을 제공하고 있다. 나이키는 '로블록스'에 '나이키랜드'를 구축해 사용자들이 가상공간에서 브랜드를 체험하고 게임을 즐길 수 있도록 했다. 구찌는 메타버스에서 한정판 가상 아이템을 판매하며, 브랜드 경험을 확장하고 있다.

H&M은 '세컨드 라이프' 플랫폼에 가상 매장을 열어, 고객들이 아바타를 통해 제품을 체험하고 구매할 수 있는 환경을 제공하고 있다. BMW는 '호라이즌 워크룸'을 활용해 가상 회의와 협업을 통해 제품 개발과 디자인을 진행하고 있다. 이러한 사례들은 해외 기업들이 메타버스를 통해 비즈니스 모델을 혁신하고, 고객과의 새로운 접점을 만들어가고 있음을 보여준다.

라. 헬스케어

헬스케어 분야에서도 메타버스를 통해 진단, 치료, 교육 등 다양한 분야에서 혁신적인 변화가 나타나고 있다. 국내에서는 중앙대학교 광명병원은 메타버스를 활용한 가상 병원을 구축하고, 진료와 상담을 가상 환경에서 시연하며 의료 혁신의 가능성을 제시했다. 이 플랫폼은 환자가 의료진과 가상공간에서 소통하며 편리하게 진료를 받을 수 있는 환경을 조성한다. 분당서울대병원은 메타버스 기반의 수술 시뮬레이션 시스템을 통해 의료진이 가상공간에서 복잡한 수술을 연습하고, 교육할 수 있도록 하고 있다. 이는 의료 사고를 줄이고, 보다 안전한 진료를 제공하는 데 기여한다. 룩시드랩스는 VR 기기를 통해 뇌파와 시선 추적 데이터를 분석하여 정신질환을 조기에 진단하는 기술을 개발했으며, 노인 치매 예방에도 활용되고 있다.

라운메타는 메타버스 기반 간호 실습 콘텐츠를 개발해 간호사들이 실제 의료 상황을 가상으로 체험할 수 있도록 지원하고 있다. 이 기술은 간호사의 역량 강화와 스트레스 관리에 도움을 준다. 아울러 국내 스타트업들은 메타버스를 활용한 재활치료 프로그램을 선보이고 있다. 이 프로그램은 환자가 가상 환경에서 운동을 수행하며 물리치료를 받을 수 있는 시스템을 포함하고 있다. 일부 병원에서는 가상현실 기반의 정신건강 치료를 도입하여 환자들이 메타버스 공간에서 안정과 심리적 치유를 경험하도록 돕고 있다.

해외에서는 메타버스를 활용한 헬스케어 기술이 더욱 정교하고 광범위하게 적용되고 있다. 미국의 존스홉킨스 병원은 수술 시뮬레이션을 메타버스를 통해 구현하여 의료진이 수술 전에 복잡한 수술을 연습할 수 있는 환경을 제공한다. 이를 통해 수술 성공률을 높이고, 환자의 위험을 줄이는 데 기여하고 있다. 영국의 NHS는 메타버스를 활용한 정신건강 치료 프로그램을 도입해 환자들이 가상 환경에서 스트레스 관리와 치료를 받을 수 있도록 지원하고 있다. 프랑스의 아소부는 메타버스 플랫폼을 통해 물리치료와 재활 훈련을 가상으로 제공하며, 환자들이 치료 과정에서 더욱 몰입할 수 있도록 돕고 있다.

일본의 후지쓰는 원격 의료 상담 서비스를 메타버스 환경에서 제공하고, 환자와 의료진 간의 상호작용을 개선하고 있다. 캐나다의 메디컬 리얼리티스는 의료 교육 플랫폼을 개발해 전 세계 의료진이 메타버스를 통해 실시간으로 교육을 받을 수 있는 환경을 제공한다. 이 플랫폼은 실제 사례 기반의 학습을 지원하며, 의료진의 역량 강화를 돕고 있다. 또한, 독일의 의료 기술 회사는 메타버스를 활용해 암 치료 계획을 시뮬레이션하여 보다 정밀하

고 효과적인 치료 방안을 설계하고 있다. 해외 헬스케어 업계는 메타버스를 통해 치료의 질을 높이고, 교육과 진단의 새로운 표준을 제시하고 있다.

마. 공공서비스

공공서비스 영역도 예외가 아니다. 국내에서는 메타버스를 활용하여 공공서비스를 혁신하려는 시도가 활발히 이루어지고 있다. 서울시는 ‘메타버스 서울’ 프로젝트를 통해 시민들이 가상공간에서 시정에 참여하거나 행정 서비스를 이용할 수 있는 플랫폼을 구축했다. 이를 통해 시민들은 민원 처리, 세금 납부, 시정회의 참석 등을 가상 환경에서 편리하게 수행할 수 있다. 경상북도는 메타버스 플랫폼을 활용하여 지역의 관광지를 가상으로 체험할 수 있는 서비스를 제공하고 있다. 예를 들어, 안동 하회마을이나 경주의 역사적 장소를 가상 환경에서 탐방할 수 있도록 구현해 지역 관광을 활성화하고 있다.

부산시는 메타버스 기반 스마트시티 플랫폼을 도입하여, 시민들이 가상공간에서 도시 계획과 발전 방향에 대한 의견을 제시할 수 있는 환경을 마련했다. 광주광역시도 지역 예술가들의 작품을 가상 전시회 형태로 선보이며, 시민들과의 문화적 소통을 강화하고 있다. 대전시는 메타버스 플랫폼을 통해 공공 정보를 가상공간에서 제공하며, 시민들과의 소통을 증진시키고 있다. 이러한 사례들은 메타버스가 공공서비스의 접근성을 높이고, 시민 참여를 확대하는 데 기여하고 있음을 보여준다.

싱가포르는 ‘버추얼 싱가포르’ 프로젝트를 통해 도시 전체를 3D로 구현해 도시 관리와 계획에 활용하고 있다. 이 플랫폼은 정부, 기업, 연구기관이 데이터를 공유하고 협업하는데 사용되며, 도시 문제를 예측하고 해결하는 데 기여하고 있다. 두바이는 메타버스 기반 스마트시티 플랫폼을 구축하여, 시민들이 가상공간에서 민원 처리, 행정 서비스 이용 등 다양한 공공서비스를 받을 수 있도록 하고 있다.

핀란드는 가상 대사관을 메타버스를 통해 운영하며, 해외에 거주하는 자국민들에게 영사 서비스를 제공하고 있다. 예를 들어, 여권 갱신이나 비자 신청 등을 가상공간에서 처리할 수 있다. 미국 뉴욕시는 메타버스 플랫폼을 활용하여 시민들과 소통하고, 공공 정보를 시각적으로 제공하는 서비스를 운영 중이다. 일본 도쿄도는 가상 관광 서비스를 도입해 전 세계 사용자들이 메타버스를 통해 도쿄의 주요 관광지를 체험할 수 있도록 지원하고 있다. 이러한 사례들은 메타버스가 공공서비스에서의 효율성과 접근성을 높이며, 글로벌 시민 참여를 촉진하는 데 중요한 도구로 자리 잡고 있음을 보여준다.

〈표 5-2〉 분야별 메타버스 활용 사례

분야	주요 활용 사례	국내 사례	해외 사례
교육	가상 학습 환경 제공, 시각적 학습 강화, 비용 효율성, 글로벌 협업 가능	메타버스 캠퍼스 투어, 입학식 (이프랜드), 가상 유적지 탐방, 디지털 인재 양성 프로그램	zSpace 가상 실험실 (미국), 메타버스 의학 교육 (영국 Imperial College London), GidiX 언어 학습 (중국), 가상 농업 훈련 (호주)
엔터테인먼트	가상 콘서트, 팬-아티스트 소통, 게임 및 가상공간 활용	제페토(K-pop 협업), 이프랜드(가상 회의), 위버스(글로벌 팬 소통), 디어유 버블(1:1 소통), 유니버스(콘텐츠 제공)	로블록스 가상 콘서트 (미국 Travis Scott), iHeartLand 팟캐스트 (미국), Fortnite 가상 이벤트
비즈니스 및 쇼핑	가상 매장 및 모델, 브랜드 체험, 가상 오피스	롯데홈쇼핑(루시), GS리테일(가상 점포), 현대차(가상 전시장), 신세계(가상 매장), LG유플러스(가상 오피스)	월마트 가상 매장(로블록스), 나이키랜드(로블록스), 구찌(가상 아이템), H&M(세컨드 라이프), BMW(호라이즌 워크룸)
헬스케어	진단 및 치료 시뮬레이션, 재활 치료, 정신 건강 관리	중앙대병원(가상 병원), 분당서울대병원(수술 시뮬레이션), 룩시드랩스(VR 기반 정신질환 진단), 재활치료 프로그램	존스홉킨스 병원(수술 연습), NHS 정신 건강 치료(영국), 아소부(재활 훈련, 프랑스), 후지쓰 원격 의료(일본), 메디컬 리얼리티스 의료 교육(캐나다), 암 치료 시뮬레이션(독일)
공공 서비스	행정 서비스, 스마트시티, 가상 관광	메타버스 서울(민원, 세금), 경북 가상 관광(안동 하회마을), 부산 스마트시티, 광주 가상 전시회, 대전 공공정보 제공	버추얼 싱가포르(도시 계획), 두바이 스마트시티, 핀란드 가상 대사관, 뉴욕 공공 정보 플랫폼, 도쿄 가상 관광 서비스

자료: 연구진 작성

제3절 메타버스 이용자 보호 기본원칙 검토

1. 메타버스 이용자 보호 기본원칙 검토

방송통신위원회는 메타버스 이용자들이 안전하고 편리하게 서비스를 이용할 수 있도록 ‘메타버스 이용자 보호 기본원칙’을 마련하여 2023년 11월 29일에 발표했다. 이 기본원칙은 메타버스 플랫폼 서비스 제공 사업자에게 권고되는 자율규범으로, 아동·청소년 및 개인 프라이버시 보호 등 이용자 보호에 대한 기준을 제시하고 있다. 주요 사업자들도 이 기본원칙 도입에 동의하였으며, 향후 약관 및 서비스 운영 규정 등에 해당 사항을 반영할 예정이다. 주요 내용은 다음과 같다.

〔그림 5-2〕 메타버스 이용자 보호 기본원칙



자료: 저자 재구성

1) 공동체 가치 형성

“가상주체 등 매개체를 통한 소통과 교류가 자아실현과 표현의 자유를 최대한 보장하되, 상호 존중과 배려를 기초로 보편타당한 공동체 가치를 형성하고 보호하도록 한다.”

메타버스는 가상공간에서 이용자가 자유롭게 자신을 표현하고 교류할 수 있는 환경을 제공한다. 이러한 환경은 현실 세계에서 억압되거나 제한되었던 자아실현의 기회를 확대하고, 창의적이고 혁신적인 활동을 가능하게 한다. 이를 위해 메타버스 플랫폼은 아바타, 가상 재화, 콘텐츠 제작 도구 등 이용자가 자신을 나타내고 소통할 수 있는 다양한 수단을

제공해야 한다. 특히 표현의 자유를 최대한 보장함으로써 메타버스가 다양한 아이디어와 문화를 공유하는 장으로 기능하도록 해야 한다.

그러나 자아실현과 표현의 자유는 다른 이용자들의 권리를 침해하지 않는 범위에서 이루어져야 한다. 이를 위해 플랫폼은 혐오 발언, 괴롭힘, 성적 폭력과 같은 부적절한 행위를 방지하기 위한 정책과 기술적 조치를 마련해야 한다. 또한, 이용자 간 상호 존중과 배려의 문화가 조성될 수 있도록 커뮤니티 가이드라인을 명확히 제시하고 이를 준수하도록 독려해야 한다. 이는 메타버스가 단순히 개인적인 공간을 넘어 공동체 가치를 형성하고 사회적 연대를 강화하는 공간으로 발전하는 데 필수적이다.

2) 이용자 참여 보장

“이용자가 제품·서비스의 운용 원리·정책을 이해할 수 있도록 하고, 이용자 권익에 관한 주요 사항에 대해 의견을 제시할 수 있는 적정 절차를 갖추어야 한다.”

메타버스 플랫폼의 작동 원리와 정책은 투명해야 하며, 이용자들이 이를 명확히 이해할 수 있어야 한다. 이는 플랫폼이 신뢰를 구축하는 데 핵심적이다. 예를 들어, 데이터 수집 및 처리 방식을 명시적으로 설명하고, 사용자가 자신의 데이터를 어떻게 관리할 수 있는지에 대한 정보를 제공해야 한다. 또한, 서비스 약관 및 이용 규정을 평이한 언어로 작성하여 모든 이용자가 쉽게 이해할 수 있도록 해야 한다.

이용자의 권익을 보호하기 위해, 이용자들이 정책이나 서비스 운영에 관한 의견을 제시할 수 있는 절차를 마련해야 한다. 이를 통해 이용자와 플랫폼 간의 상호작용을 강화하고, 이용자의 피드백을 바탕으로 정책을 개선할 수 있다. 설문조사, 이용자 포럼, 신고 시스템 등 다양한 채널을 활용하여 이용자 의견을 수렴하는 것이 바람직하다. 이는 플랫폼의 지속 가능한 성장을 도모하고, 이용자가 플랫폼의 운영 과정에 참여할 수 있는 기회를 제공함으로써 커뮤니티의 신뢰를 높이는 데 기여한다.

3) 이용자 갈등 조정

“이용자 간에 발생하는 갈등을 합리적으로 조정하고, 기본적인 소통과 교류의 방식에 관하여 이용자에게 선택권을 부여하기 위해 노력한다.”

메타버스는 다양한 배경과 관점을 가진 이용자들이 상호작용하는 공간으로, 갈등이 발생할 가능성이 높다. 이러한 갈등을 합리적으로 조정하기 위해 플랫폼은 명확한 분쟁 해

결 절차와 중재 시스템을 마련해야 한다. 예를 들어, 신고 기능과 분쟁 중재 기구를 통해 이용자가 겪는 문제를 신속하고 공정하게 해결할 수 있도록 해야 한다. 또한, 플랫폼은 갈등 발생 가능성을 줄이기 위해 초기 단계에서부터 커뮤니티 규칙을 명확히 설정하고, 이를 적극적으로 알릴 필요가 있다.

소통과 교류 방식에 대한 이용자의 선택권은 갈등 예방과 플랫폼 참여를 높이는 데 중요한 역할을 한다. 이용자가 자신의 소통 방식을 설정할 수 있도록 프라이버시 옵션, 블록 기능, 대화 제한 설정 등을 제공해야 한다. 예를 들어, 아바타 간의 거리 유지 설정이나 특정 이용자와의 상호작용을 제한하는 기능은 이용자들이 불편한 상황을 회피하고 더 나은 경험을 할 수 있도록 돕는다. 선택권을 부여함으로써 이용자는 자신에게 적합한 소통 환경을 구축할 수 있다.

4) 공정한 경제활동 보호

“메타버스에서 유통되는 디지털 재화·서비스 등의 거래가 공정하고 합리적이며 비차별적인 조건으로 이루어지도록 하고, 지식재산권의 침해를 방지해야 한다.”

메타버스에서 디지털 재화 및 서비스는 경제 활동의 핵심 요소로 자리 잡고 있다. 이들 거래가 공정하고 투명하게 이루어지도록 플랫폼은 명확한 거래 규칙과 안전장치를 마련해야 한다. 이를 통해 사기나 가격 담합과 같은 문제를 예방할 수 있다. 또한, 거래 중 발생할 수 있는 분쟁 해결을 위해 신뢰할 수 있는 제3자 중재 시스템을 도입하는 것도 필요하다. 이용자가 거래 조건을 명확히 이해하고 동의한 상태에서 거래가 이루어지도록 사전 고지 시스템을 구축해야 한다.

디지털 재화와 관련된 지식재산권 보호는 플랫폼의 신뢰성을 강화하는 데 중요한 요소이다. 크리에이터의 콘텐츠가 불법 복제되거나 도용되지 않도록 기술적 보호 장치를 마련하고, 지식재산권 침해 사례를 신속히 처리할 수 있는 절차를 제공해야 한다. 예를 들어, NFT나 디지털 아트와 같은 콘텐츠는 블록체인 기술을 통해 소유권을 명확히 할 수 있다. 이러한 노력은 창작자와 소비자 모두에게 공정한 경제 생태계를 제공한다.

5) 데이터 통제권 보장

“이용자가 본인과 매개체의 데이터를 손쉽게 효과적으로 관리·통제·활용할 수 있는 권리를 보장해야 한다.”

메타버스는 이용자의 개인정보와 활동 데이터를 대규모로 수집하고 활용한다. 따라서 이용자가 자신의 데이터에 대한 소유권과 통제권을 가지는 것은 필수적이다. 플랫폼은 데이터를 확인, 수정, 삭제할 수 있는 도구를 제공하고, 데이터 활용에 대한 동의를 명확히 받을 의무가 있다. 이를 통해 이용자는 자신의 데이터가 어떻게 사용되는지에 대해 투명하게 알 수 있다.

데이터 통제 권리는 이용자들에게 데이터 활용의 이점을 제공하면서도 프라이버시를 보호할 수 있도록 한다. 예를 들어, 이용자가 자신의 활동 데이터를 분석하여 개인화된 서비스 추천을 받을 수 있으나, 동시에 필요에 따라 데이터를 삭제하거나 공유를 제한할 수 있어야 한다. 이러한 데이터 관리 권리는 이용자들에게 안전한 이용 환경을 제공할 뿐만 아니라, 메타버스 서비스에 대한 신뢰를 구축하는 데 중요한 역할을 한다.

6) 지속가능발전 지향

“제품·서비스의 지속가능발전목표를 위한 노력을 공개하고, 메타버스가 이용자의 신체적·정신적 건강 및 사회·문화·환경·경제 등에 미치는 장기적인 영향을 연구하기 위해 노력한다.”

메타버스는 단순한 가상공간을 넘어, 사회적, 문화적, 경제적 영향을 끼치는 플랫폼으로 발전하고 있다. 이에 따라 메타버스 서비스 제공자는 지속가능발전목표(SDGs)를 반영하여 서비스 설계를 해야 한다. 예를 들어, 탄소 배출 감소를 위한 에너지 효율적인 서버 운영, 환경 친화적인 디지털 자산 거래를 도입하는 등의 노력이 포함된다. 이러한 투명성을 보장하기 위해 플랫폼은 지속가능발전과 관련된 노력을 주기적으로 공개하고, 이용자와 이해관계자들에게 이를 알릴 책임이 있다.

메타버스가 이용자의 신체적, 정신적 건강에 미치는 영향을 고려하는 것은 매우 중요하다. 지나친 메타버스 사용으로 인한 중독, 사회적 고립, 정서적 불안 등 부작용을 방지하기 위해 플랫폼은 이용자 행동 데이터를 분석하고 이를 기반으로 경고 및 휴식 권장 기능을 도입할 수 있다. 또한, 메타버스의 사회적 및 경제적 영향을 분석하여 긍정적인 방향으로 발전할 수 있도록 연구와 투자를 확대해야 한다. 이를 통해 메타버스는 이용자 개인뿐만 아니라 전 세계적으로 지속 가능한 발전에 기여할 수 있다.

2. 이용자 보호 기본원칙 검토

메타버스 이용자 보호 기본원칙은 메타버스에서 이용자의 자아실현과 표현의 자유를 보장하면서도, 상호 존중과 배려를 통해 보편적 공동체 가치를 보호하는 것을 목표로 한다. 혐오와 차별, 성추행, 불법 콘텐츠 유통을 금지하며, 데이터 수집과 콘텐츠 제한, 수익 배분 구조를 투명하게 공개하여 이용자 참여를 보장한다. 또한, 이용자 간의 갈등은 적정한 신고 절차와 대체적 분쟁 해결 절차를 통해 조정하고, 개인적인 소통 방식과 매개체 보호를 선택할 권리를 부여한다. 공정한 경제활동 보호를 위해 디지털 재화와 서비스 거래에서 공정성과 비차별성을 보장하고, NFT 거래와 저작권 보호를 포함한 지식재산권 침해 방지 조치를 명시한다. 데이터 통제권은 이용자와 매개체의 데이터를 쉽게 관리하고 활용할 수 있는 방식으로 보장하며, 메타버스가 사회적, 경제적, 환경적 지속 가능성을 유지하도록 노력한다.

메타버스의 지속 가능성을 위해 이용자 건강, 사회·문화적 영향 등을 장기적으로 연구하고, 지속 가능 발전 목표(SDGs)에 부합하는 서비스 제공 방안을 공개한다. 이는 메타버스가 개인과 사회에 미치는 긍정적 효과를 극대화하고, 부정적 영향을 최소화하기 위한 체계적 접근을 포함한다. 특히, 공정한 거래 관행과 이용자 데이터 보호는 메타버스의 신뢰 구축과 지속 가능성의 핵심으로 간주된다.

메타버스 사업자들이 기본원칙을 충실히 이행하는지 평가하기 위해, 주요 플랫폼을 대상으로 정책 및 운영 실태를 점검한다. 조사 범위는 데이터 수집·활용 방식, 콘텐츠 제한 및 추천 알고리즘의 공정성, 이용자 데이터 보호와 관리 수준 등을 포함하며, 이를 위해 문서 검토, 설문 조사, 인터뷰를 진행한다. 또한, 이용자 간 갈등 관리 체계와 불법 콘텐츠 대응 방안을 실증적으로 점검하기 위해, 신고 처리 사례 및 대체적 분쟁 해결 절차의 투명성을 분석한다. 공정한 경제활동을 위해 NFT 거래와 관련한 명확한 권리 명시, 불공정 거래 방지 조치를 중점적으로 조사한다.

특히 중점적으로 점검할 원칙은 투명성과 데이터 통제권이다. 투명성은 데이터 활용 방식과 수익 배분 구조의 명료성을 통해 이용자 신뢰를 구축하는 데 필수적이다. 데이터 통제권은 이용자 개인 정보와 생체 데이터 보호 및 활용 권리를 직접적으로 다루는 만큼, 이용자 권익과 사생활 보호를 위한 가장 중요한 원칙으로 평가된다. 이를 기반으로 메타버

스 플랫폼이 이용자 친화적이고 책임 있는 디지털 환경을 조성하고 있는지를 종합적으로 평가하고자 하였다.

제 4 절 메타버스 사업자의 이용자 보호 실태 조사

1. 사업자 A

가. 공동체 가치 형성

사업자 A의 ‘공동체 가치 형성’ 항목에 대한 설문조사 결과는 모든 항목에서 ‘매우 그렇다(5/5)’라고 매우 긍정적으로 응답하였다. 구체적으로, 부적절한 콘텐츠, 미성년자에게 유해한 콘텐츠, 성적 문제를 일으킬 수 있는 콘텐츠, 자살 및 자해 관련 콘텐츠, 폭력 관련 콘텐츠, 위험한 행위 및 테러리즘 관련 콘텐츠, 혐오 행위 관련 콘텐츠, 괴롭힘 및 따돌림 관련 콘텐츠, 불법 행위 및 규제 물품 관련 콘텐츠 등 다양한 유형의 유해 콘텐츠를 철저히 제한하고 있으며, 플랫폼의 안정성을 해치는 콘텐츠와 개인정보 무단 송출을 제한하는 데도 매우 신경 쓰고 있다고 응답하였다.

〈표 5-3〉 공동체 가치 형성 관련 조사 내용-사업자 A

번호	항목	응답결과
A1	부적절한 콘텐츠 제한	매우 그렇다(5/5)
A1-1	미성년자 유해 콘텐츠 제한	매우 그렇다(5/5)
A1-2	성적 문제 콘텐츠 제한	매우 그렇다(5/5)
A1-3	자살 및 자해 관련 콘텐츠 제한	매우 그렇다(5/5)
A1-4	폭력 관련 콘텐츠 제한	매우 그렇다(5/5)
A1-5	위험/테러리즘/극단주의 콘텐츠 제한	매우 그렇다(5/5)
A1-6	혐오 행위 관련 콘텐츠 제한	매우 그렇다(5/5)
A1-7	괴롭힘/따돌림 관련 콘텐츠 제한	매우 그렇다(5/5)
A1-8	불법 행위/규제 물품 관련 콘텐츠 제한	매우 그렇다(5/5)
A2	플랫폼 안정성 방해 콘텐츠 제한	매우 그렇다(5/5)
A3	개인정보 무단 송출 제한	매우 그렇다(5/5)

다만, 이는 자체 평가결과이므로 구체적으로 유해 콘텐츠를 어떤 기준으로 판별하는지, 어떤 방법으로 유해 콘텐츠를 제한하는지, 제한하는 방법이 얼마나 효과적인지, 새롭게 만들거나 보완하거나 규제기준은 없는지, 새로운 형식과 내용의 유해 콘텐츠는 없는지, 사업자의 실제 이용자들의 경험과 일치하는지 등에 대한 추가적인 검증 및 지속적인 모니터링 필요할 것으로 보인다.

나. 이용자 참여 보장

사업자 A의 ‘이용자 참여 보장’ 항목에 대한 설문조사 결과는 대체로 매우 긍정적이었다. 구체적으로, 사용자 콘텐츠의 정의와 범위, 저작권 및 지식재산권 소유 및 귀속, 노출 범위와 조정 방법, 관련 문제 및 분쟁에 대한 책임 소지, 서비스 내외에서의 활용 방식, 그리고 서비스 이용 시 발생 가능한 법적 문제에 대한 유의사항 등을 매우 명확하게 제시하고 있다고 응답했다.

〈표 5-4〉 이용자 참여 보장 관련 조사 내용-사업자 A

번호	항목	결과
B1	사용자 콘텐츠에 대한 정의와 범위 제시	매우 그렇다(5/5)
B2	사용자 콘텐츠의 저작권 및 지식재산권 소유 및 귀속 제시	매우 그렇다(5/5)
B3	사용자 콘텐츠의 노출 범위와 조정 방법 제시	매우 그렇다(5/5)
B4	사용자 콘텐츠 관련 문제 제기 및 분쟁에 대한 책임 소지 제시	매우 그렇다(5/5)
B5	서비스상 사용자 콘텐츠 활용 방식 및 외부 활용 방식 제시	매우 그렇다(5/5)
B6	서비스 이용 시 발생 가능한 법적 문제에 대한 유의사항 제시	매우 그렇다(5/5)
B7	사용자 콘텐츠를 통한 2차 생산물 제작 가능 범위와 규제 제시	잘 모른다.

다만, ‘사용자 콘텐츠를 통한 2차 생산물 제작 가능 범위와 규제 제시’에 대해서는 ‘잘 모른다’라고 응답하였다. 메타버스에서 ‘사용자 콘텐츠를 통한 2차 생산물’은 사용자가 직접 만든 콘텐츠를 기반으로 새롭게 창작되거나 변형된 콘텐츠를 의미한다. 예를 들어, 다른 사용자가 만든 아바타 의상이나 액세서리를 구매하여 자신의 아바타에 적용하는 아바타 커스터마이징, 다른 사용자가 만든 가상 건물이나 환경을 구매하여 자신의 가상공간에 배치하거나 수정하는 가상공간 제작, 다른 사용자가 만든 게임 맵을 수정하거나 확장하여

새로운 게임을 만드는 게임 맵 제작, 다른 사용자가 만든 디지털 아트를 구매하여 자신의 가상 갤러리에 전시하거나 수정하는 디지털 아트워크 등이 2차 생산물이다. 2차 생산물 시장은 지속적으로 성장하고 있으므로, 이에 대한 명확한 규정을 만들어 제시할 필요가 있다.

다. 이용자 갈등 조정

‘이용자 갈등 조정’ 항목에 대한 사업자 A의 설문결과는 전반적으로 매우 긍정적이었다. 이용자 간 갈등이나 사용자 콘텐츠 관련 이의를 접수하고 처리할 수 있는 제도를 매우 잘 마련하고 있으며(5/5), 이용자가 다른 이용자의 콘텐츠나 활동이 규정에 위반된다고 판단할 시 신고할 수 있는 방법을 매우 효과적으로 제공하고 있다고 응답했다(5/5). 신고 방법은 이용자가 이해하기 쉽게 안내되어 있으며(5/5), 구체적으로 ‘앱에서 신고가 가능’하다고 답했다. 또한 사용자 콘텐츠와 관련한 이용자의 궁금증이나 불만 사항, 분쟁을 해결할 수 있는 고객대응 창구를 매우 잘 운영하고 있고(5/5), 이용자가 고객대응 창구를 이용할 수 있는 상황에 대한 안내가 매우 명확하며(5/5), 이용자가 스스로 문제를 해결할 수 있는 기능도 제공하고 있다(5/5)고 답했다.

〈표 5-5〉 이용자 갈등 조정 관련 조사 내용－사업자 A

번호	항목	결과
C1	이용자 간 갈등 혹은 사용자 콘텐츠 관련한 이의를 접수, 처리할 수 있는 제도 마련	매우 그렇다(5/5)
C2	이용자가 다른 이용자의 콘텐츠나 활동이 규정에 위반된다고 판단할 시, 신고할 수 있는 방법 제공	매우 그렇다(5/5)
C2-1	신고 방법은 이용자가 이해하기 쉽게 안내	매우 그렇다(5/5)
C2-2	구체적인 신고 방법	앱에서 신고 가능
C3	사용자 콘텐츠와 관련한 이용자의 궁금증이나 불만 사항, 분쟁을 해결할 수 있는 고객대응 창구 운영	매우 그렇다(5/5)
C3-1	이용자가 고객대응 창구를 이용할 수 있는 상황에 대한 안내	매우 그렇다(5/5)
C3-2	이용자가 스스로 문제 해결을 해볼 수 있는 메뉴 혹은 기능 존재	매우 그렇다(5/5)
C3-3	다양한 루트의 고객대응 창구 운영	그런 편이다(4/5)
C3-4	구체적인 고객대응 창구	설정가능, Help Center

다양한 루트의 고객대응 창구를 운영하고 있는지 묻는 항목에는 그런 편이다(4/5)로 응답하여 다른 항목에 비해 상대적으로 낮았으나, 그럼에도 긍정적으로 답하였다. 구체적인 고객 대응 창구는 이용자가 설정할 수 있는 기능을 제공하고, Help Center를 통해 이용자들이 문의나 불만을 제기할 수 있다고 응답했다. 전반적으로 사업자 A는 이용자 갈등 조정과 관련하여 잘 대응하고 있다고 인식하고 있는 것으로 나타났다.

그럼에도 이는 자체 평가 결과이므로 이용자 간 갈등과 조정 및 신고 처리 제도에 대한 구체적인 내용과 고객대응 체계를 별도로 파악하고, 이러한 인식이 실제로 이용자들의 경험과 얼마나 일치하는지 점검할 필요가 있을 것으로 보인다. 또한 구체적인 신고방법을 묻는 질문에 ‘앱에서 신고 가능’이라는 응답은 구체성이 부족하고, 신고 절차의 용이성, 처리 시간, 피드백 제공 등에 대한 상세 정보가 없어 실효성을 판단하기 어렵다. 이용자가 스스로 문제를 해결할 수 있는 구체적인 기능에 대해서도 파악해 볼 필요가 있을 것이다. 마지막으로, 이용자 간 갈등 해결을 위한 구체적인 중재 절차나 정책이 있는지 점검하고, Help Center이외에 구체적인 고객대응 창구 마련도 제안할 필요가 있다.

라. 공정한 경제활동 보호

사업자 A는 ‘공정한 경제활동 보호’와 관련된 모든 항목에 대해 ‘아직 준비 중’이라고 답변하였다. 이는 아직 가상 거래의 정의, 가치, 허용 범위, 저작권 및 상표권 침해 등에 대한 명확한 안내가 없고, 문제 발생 시 도움을 청할 수 있는 창구가 없다는 것은 이용자들이 거래 시 발생할 수 있는 위험에 무방비로 노출될 수 있고 불법적인 거래가 발생할 가능성이 있음을 의미한다. 또한 법적 분쟁이 발생할 경우 적절한 해결책을 제시하기 어렵고, 메타 코리아와 이용자 모두에게 법적 위험을 초래할 수 있다.

〈표 5-6〉 공정한 경제활동 보호 관련 조사 내용—사업자 A

번호	항목	결과
D1	가상 콘텐츠 및 아이템 거래의 정의에 대한 명확한 안내	준비 중
D2	가상 콘텐츠 및 아이템 거래가 실제 세계에서 지니는 가치에 대한 명확한 안내	준비 중
D3	가상 콘텐츠 및 아이템 거래가 허용되는 상황, 불허되는 상황에 대한 명확한 안내	준비 중
D4	가상 콘텐츠 및 아이템 거래 시 저작권 및 상표권 침해 소지가 있는 경우에 대한 명확한 안내	준비 중
D5	가상 콘텐츠 및 아이템 거래와 관련하여 저작권 및 상표권 문제 발생 시 도움을 청할 수 있는 창구 운영	준비 중
D5-1	도움을 청할 수 있는 창구를 이용자가 쉽게 찾을 수 있도록 안내	준비 중

이러한 중요한 정보들이 제공되지 않는다는 것은 사업자 A가 책임을 회피하고 운영 방식이 투명하지 않다는 인식을 줄 수 있다. 또한 기본적인 거래 규칙과 보호 장치가 마련되지 않은 상태에서 서비스를 운영하고 있다는 것은 서비스의 완성도와 안정성에 의문을 제기하게 만들 수 있다. 그 결과 이용자의 신뢰를 잃고, 법적, 윤리적 문제에 직면할 가능성이 높기 때문에 안전한 거래 환경 조성을 위한 방안을 시급히 마련해야 할 것으로 보인다.

마. 데이터통제권 보장

데이터통제권 보장에 대한 사업자 A의 자체 평가는 전반적으로 매우 긍정적이었다. 이용자가 직접 제공하는 정보와 서비스 사용 시 자동으로 수집되는 정보에 대한 명확한 안내가 잘 이루어지고 있으며, 이용자 개인정보 처리 목적에 대해서도 명확하게 안내하고 있다고 응답했다. 또한, 이용자가 자신의 개인정보를 열람하고 수정할 수 있는 방법에 대한 안내 역시 매우 명확하게 제공되고 있고, 개인정보 관련 이용자 권리에 대한 안내도 매우 잘 이루어져 있다고 답했다. 개인정보 이용자 권리 행사 및 문의 시 연락 방법에 대해서는 ‘그런 편이다’(4/5)로 다른 항목에 비해 약간의 개선의 여지가 있는 것으로 평가되었다.

〈표 5-7〉 데이터통제권 보장 관련 조사 내용 - 사업자 A

번호	항목	결과
E1	이용자가 직접 제공하는 정보에 대한 명확한 안내	매우 그렇다(5/5)
E2	서비스 사용 시 자동으로 수집되는 정보에 대한 명확한 안내	매우 그렇다(5/5)
E3	이용자 개인정보 처리 목적에 대한 명확한 안내	매우 그렇다(5/5)
E4	이용자의 개인정보 열람 및 수정 방법에 대한 쉽고 명확한 안내	매우 그렇다(5/5)
E5	개인정보 관련 이용자 권리에 대한 명확한 안내	매우 그렇다(5/5)
E5-1	개인정보 이용자 권리 행사 및 문의 시 연락 방법에 대한 쉽고 명확한 안내	그런 편이다(4/5)

그럼에도 이용자들의 데이터통제권에 대한 사업자의 대응 및 조치는 특히 실제로 이용자 입장에서 평가해볼 필요가 있다. 민감한 개인정보가 구체적으로 어디까지 수집되고 어떤 목적으로 어디에 사용되는지에 대한 명확한 안내가 제공되는지 철저하게 이용자 입장에서 평가할 필요가 있다는 것이다. 이를 위해 외부 평가나 사용자 피드백의 과정을 반드시 거쳐야 할 것으로 보인다. 이와 함께 사업자는 이용자들의 데이터통제권과 관련된 정책에 대해 지속적으로 확인하여 변화하는 법적 요구사항과 기술 발전에 대응할 수 있도록 해야 할 것이다.

바. 지속가능발전 지향

지속가능발전을 위한 노력을 해 오고 있는지를 조사한 결과, 사업자 A는 전반적으로 긍정적인 답변을 하였다. 구체적으로, 이용자에게 미치는 영향을 평가하고 대응하기 위한 노력에 대해 ‘매우 그렇다’(5/5)라고 응답하면서 구체적인 노력 사례로는 AI와 휴먼리뷰를 통한 콘텐츠 검수, Help Center 운영 등을 제시했다. 이용자에게 미치는 영향에 대한 평가와 대응의 정례적 실시, 내부구성원 공개, 그리고 이용자 피드백 수집에 대해서는 모두 ‘그런 편이다’(4/5)라고 응답했다. 조사 결과에 따르면, 사업자 A는 이용자에게 미치는 영향을 지속적으로 모니터링하고, 그 결과를 내부적으로 공유하며, 이용자의 의견을 수렴하는 데 어느 정도 노력을 기울이고 있다고 스스로 평가하였다.

〈표 5-8〉 지속가능발전 지향 관련 조사 내용 - 사업자 A

번호	항목	결과
F1	메타버스 서비스가 이용자에게 미치는 영향을 평가하고 대응하기 위한 노력	매우 그렇다(5/5)
F1-1	구체적인 노력 사례	AI와 휴먼리뷰를 통한 콘텐츠 검수, Help Center 운영 등
F1-2	이용자에게 미치는 영향에 대한 평가와 대응의 정례적 실시	그런 편이다(4/5)
F1-3	이용자에게 미치는 영향에 대한 평가와 대응의 내부구성원 공개	그런 편이다(4/5)
F1-4	이용자에게 미치는 영향에 대한 평가와 대응에 대한 이용자 피드백 수집	그런 편이다(4/5)

이러한 자체 응답 조사에 추가하여 이용자에 미치는 영향에 대한 평가 및 대응을 위해 구체적으로 어떤 노력을 기울이는지, 평가와 대응을 얼마나 자주 실시하는지, 평가와 대응에 대한 이용자들의 피드백을 어떤 방법으로 수집하고 이를 어떻게 활용하는지 등을 점검할 필요가 있어 보인다. 또한 이용자에게 미치는 영향을 평가하고 대응하는 노력의 일환으로 AI와 휴먼리뷰를 통한 콘텐츠 검수 및 Help Center 운영을 제시하였으나 이는 메타버스의 잠재된 영향력을 발견하고 선제적으로 대응하는 데에 부족한 방법으로, 이를 보완할 필요가 있을 것으로 판단된다.

사. 정부역할 - 서비스 제공자 지원 및 이용자 역량 강화와 보호

전반적으로 사업자 A는 이용자 보호를 위한 정부의 적극적인 역할에 대해 긍정적인 입장을 보이고 있다. 구체적으로, 서비스 제공자에게 기본원칙에 대한 실행 가이드 제공, 서비스 제공자에 대한 지원 정책 수립과 자율규제를 위한 가이드 제공, 서비스 제공자의 기본원칙 준수 유무 관리 감독 등에 동의하는 입장을 보였다. 특히 정부가 서비스 이용자의 인공지능 이용 역량을 강화하고 서비스 제공자의 이용자 보호 책무 이행을 촉구해야 한다는 데 매우 동의하는 것으로 나타났다. 사업자 입장에서는 정부의 규제 및 관리 감독이 자율성을 침해하고 성장을 억누른다고 생각할 수 있다는 우려와 달리 정부의 적극적인 역할이 필요하다는 인식하는 것으로 나타났다.

〈표 5-9〉 정부역할 관련 조사 내용—사업자 A

번호	항목	결과
G1-1	서비스 제공자에게 기본원칙에 대한 실행 가이드 제공	약간 동의한다(4/5)
G1-2	서비스 제공자에 대한 지원 정책 수립	약간 동의한다(4/5)
G1-3	서비스 제공자의 자율규제 활동을 위한 구체적인 실행 가이드 제공	약간 동의한다(4/5)
G1-4	서비스 제공자의 기본원칙 준수 유무 관리 감독	약간 동의한다(4/5)
G1-5	정부역할 관련 추가 의견	-
G2-1	서비스 이용자의 인공지능 이용 역량 강화를 위한 지원	매우 동의한다(5/5)
G2-2	서비스 제공자의 이용자 보호 책무 이행 촉구	매우 동의한다(5/5)
G2-3	정부역할 관련 추가 의견	-

2. 사업자 B

가. 공동체 가치 형성

사업자 B는 공동체 가치 형성을 위해 다양한 유해 콘텐츠와 부적절한 행위를 적극적으로 제한하고 있다고 응답하였다. 구체적으로, 미성년자 보호, 성적 문제, 자살 및 자해, 폭력, 테러리즘, 혐오 행위, 괴롭힘, 불법 행위 등과 관련된 콘텐츠 배포 및 활동을 매우 엄격하게 제한하고 있으며, 플랫폼의 안정성과 서비스의 정상적인 운영을 방해하는 콘텐츠의 배포도 강력하게 제한하고 있었다.

타인의 개인정보 무단 송출을 제한하는 것에 대해서는 ‘그런 편이다’로 응답하여, 다른 항목들에 비해 상대적으로 낮은 평가를 하여 추가적인 개선의 여지가 있음이 보였으나, 전반적으로 공동체 가치 형성을 위한 다양한 제한 정책을 적극적으로 실시하고 있으며, 이를 통해 안전하고 건전한 메타버스 환경을 조성하려는 노력을 기울이고 있음을 보여주었다.

〈표 5-10〉 공동체 가치 형성 관련 조사 내용 - 사업자 B

번호	항목	응답결과
A1	부적절한 콘텐츠 제한	매우 그렇다(5/5)
A1-1	미성년자 유해 콘텐츠 제한	매우 그렇다(5/5)
A1-2	성적 문제 콘텐츠 제한	매우 그렇다(5/5)
A1-3	자살 및 자해 관련 콘텐츠 제한	매우 그렇다(5/5)
A1-4	폭력 관련 콘텐츠 제한	매우 그렇다(5/5)
A1-5	위험/테러리즘/극단주의 콘텐츠 제한	매우 그렇다(5/5)
A1-6	혐오 행위 관련 콘텐츠 제한	매우 그렇다(5/5)
A1-7	괴롭힘/따돌림 관련 콘텐츠 제한	매우 그렇다(5/5)
A1-8	불법 행위/규제 물품 관련 콘텐츠 제한	매우 그렇다(5/5)
A2	플랫폼 안정성 방해 콘텐츠 제한	매우 그렇다(5/5)
A3	개인정보 무단 송출 제한	그런 편이다(4/5)

나. 이용자 참여 보장

이용자 참여 보장 측면에서도 대부분의 항목에서 높은 평가를 하였다. 구체적으로, 사용자 콘텐츠에 대한 정의와 범위, 저작권 및 지식재산권 관련 정보, 콘텐츠 노출 범위와 조정 방법, 문제 발생 시 책임 소지, 그리고 콘텐츠 활용 방식에 대해 매우 명확하게 제시하고 있다고 응답했다.

〈표 5-11〉 이용자 참여 보장 관련 조사 내용 - 사업자 B

번호	항목	결과
B1	사용자 콘텐츠에 대한 정의와 범위 제시	매우 그렇다(5/5)
B2	사용자 콘텐츠의 저작권 및 지식재산권 소유 및 귀속 제시	매우 그렇다(5/5)
B3	사용자 콘텐츠의 노출 범위와 조정 방법 제시	매우 그렇다(5/5)
B4	사용자 콘텐츠 관련 문제 제기 및 분쟁에 대한 책임 소지 제시	매우 그렇다(5/5)
B5	서비스상 사용자 콘텐츠 활용 방식 및 외부 활용 방식 제시	매우 그렇다(5/5)
B6	서비스 이용 시 발생 가능한 법적 문제에 대한 유의사항 제시	보통이다(3/5)
B7	사용자 콘텐츠를 통한 2차 생산물 제작 가능 범위와 규제 제시	전혀 그렇지 않다(1/5)

그러나 일부 항목에서는 개선의 여지가 있는 것으로 보인다. 서비스 이용 시 발생할 수 있는 법적 문제에 대한 유의사항 제시는 보통 수준으로 평가했으며, 특히 사용자 콘텐츠를 통한 2차 생산물 제작 가능 범위와 규제에 대한 정보 제공은 매우 부족하다고 응답하였다. 이는 서비스 이용 시 발생할 수 있는 법적 문제를 비롯하여 2차 생산물에 대한 인식과 활동이 아직 구체화되지 않았기 때문으로 보이나 구체적인 이유에 대해 탐색할 필요가 있어 보인다.

다. 이용자 갈등 조정

이용자 갈등 조정 부분에 대한 사업자 B의 응답 결과는 전반적으로 매우 긍정적으로 나타났다. 사업자 B는 이용자 간 갈등이나 사용자 콘텐츠 관련 이의를 접수하고 처리할 수 있는 제도를 매우 잘 마련하고 있고, 이용자가 다른 이용자의 콘텐츠나 활동이 규정에 위반된다고 판단할 시 신고할 수 있는 방법을 매우 효과적으로 제공하고 있다고 응답하였다. 이용자들이 필요할 때 신고할 수 있도록 서비스 쉽게 찾을 수 있는 위치에 모든 콘텐츠에 차단 및 신고 버튼을 제공하고 있다고 응답했다.

〈표 5-12〉 이용자 갈등 조정 관련 조사 내용-사업자 B

번호	항목	결과
C1	이용자 간 갈등 혹은 사용자 콘텐츠 관련한 이의를 접수, 처리할 수 있는 제도 마련	매우 그렇다(5/5)
C2	이용자가 다른 이용자의 콘텐츠나 활동이 규정에 위반된다고 판단할 시, 신고할 수 있는 방법 제공	매우 그렇다(5/5)
C2-1	신고 방법은 이용자가 이해하기 쉽게 안내	매우 그렇다(5/5)
C2-2	구체적인 신고 방법	서비스 내 모든 콘텐츠 및 사용자에게 차단 및 신고 버튼을 제공(옵션 버튼을 통해 쉽게 찾을 수 있음)
C3	사용자 콘텐츠와 관련한 이용자의 궁금증이나 불만 사항, 분쟁을 해결할 수 있는 고객대응 창구 운영	매우 그렇다(5/5)
C3-1	이용자가 고객대응 창구를 이용할 수 있는 상황에 대한 안내	매우 그렇다(5/5)

번호	항목	결과
C3-2	이용자가 스스로 문제 해결을 해볼 수 있는 메뉴 혹은 기능 존재	매우 그렇다(5/5)
C3-3	다양한 루트의 고객대응 창구 운영	매우 그렇다(5/5)
C3-4	구체적인 고객대응 창구	공식 이메일, 구글플레이스토어 내 VOC 채널, 애플앱스토어 내 VOC 채널, 윈스토어 내 VOC 채널, 114(공식 유선 채널)

더불어, 사업자 B는 사용자 콘텐츠와 관련한 이용자의 궁금증이나 불만 사항, 분쟁을 해결할 수 있는 고객대응 창구를 잘 운영하고 있고, 이용자가 고객대응 창구를 이용할 수 있는 상황에 대한 안내가 매우 명확하며, 이용자가 스스로 문제를 해결할 수 있는 기능도 제공하고 있는 것으로 나타났다. 특히 다양한 고객대응 창구를 운영하고 있는데, 공식 이메일, 구글 플레이스토어, 애플 앱스토어, 윈스토어 내 VOC 채널, 그리고 공식 유선 채널인 114 등을 통해 이용자들이 다양한 경로로 문의나 불만을 제기할 수 있도록 하고 있었다.

라. 공정한 경제활동 보호

사업자 B의 공정한 경제활동 보호 부분에 대한 자체 평가도 전반적으로 긍정적인 것으로 나타났다. 가상 콘텐츠 및 아이템 거래의 정의에 대해 매우 명확하게 안내하고 있으며 (5/5), 거래가 허용되는 상황과 불허되는 상황에 대해서도 매우 명확하게 안내하고 있다 (5/5)고 응답했다. 가상 콘텐츠 및 아이템 거래 시 저작권 및 상표권 침해 소지가 있는 경우에 대해서도 매우 명확하게 안내하고 있다(5/5)고 밝혔다. 또한 가상 콘텐츠 및 아이템 거래와 관련하여 저작권 및 상표권 문제가 발생한 경우 도움을 청할 수 있는 창구를 운영하고 있으며(5/5), 이러한 도움을 청할 수 있는 창구를 이용자가 쉽게 찾을 수 있도록 안내하고 있다(5/5)고 응답하였다.

〈표 5-13〉 공정한 경제활동 보호 관련 조사 내용-사업자 B

번호	항목	결과
D1	가상 콘텐츠 및 아이템 거래의 정의에 대한 명확한 안내	매우 그렇다(5/5)
D2	가상 콘텐츠 및 아이템 거래가 실제 세계에서 지니는 가치에 대한 명확한 안내	보통이다(3/5)
D3	가상 콘텐츠 및 아이템 거래가 허용되는 상황, 불허되는 상황에 대한 명확한 안내	매우 그렇다(5/5)
D4	가상 콘텐츠 및 아이템 거래 시 저작권 및 상표권 침해 소지가 있는 경우에 대한 명확한 안내	매우 그렇다(5/5)
D5	가상 콘텐츠 및 아이템 거래와 관련하여 저작권 및 상표권 문제 발생 시 도움을 청할 수 있는 창구 운영	매우 그렇다(5/5)
D5-1	도움을 청할 수 있는 창구를 이용자가 쉽게 찾을 수 있도록 안내	매우 그렇다(5/5)

그러나 가상 콘텐츠 및 아이템 거래가 실제 세계에서 어떠한 가치를 지니는지에 대한 안내는 보통 수준(3/5)이라고 응답하여 다른 문항에 비해 부정적 평가를 하였다. 이러한 이유는 다음과 같이 추론할 수 있다. 먼저, 가상 아이템의 실제 가치를 정확히 평가하고 설명하는 것은 매우 복잡한 과정이다. 가상 경제와 실제 경제의 연관성, 수요와 공급의 변동성, 그리고 사용자 선호도의 급격한 변화 등 다양한 요인이 영향을 미치고, 가상 재화의 법적 지위와 소유권에 대한 명확한 규정이 아직 확립되지 않아 실제 세계에서 가치 설명을 명확하게 설명하기 어려울 수 있다는 것이다. 현재 메타버스 경제가 아직 초기 단계임을 감안하면 변동성이 심한 가상 자산의 가치를 일관되게 안내하기 어려울 수 있을 것이다.

메타버스와 가상 경제에 대한 사용자들의 이해도가 다양해 모든 사용자에게 적합한 수준으로 가치를 설명하는 것이 쉽지 않다는 점도 있다. 이용자 간 기술 격차, 문화 및 경제적 차이, 메타버스 경험 수준의 차이 등에 따라 디지털 콘텐츠의 가치에 대한 인식과 평가가 다를 수 있어 범용적 설명이 어려울 수 있다는 것이다. 단계별 교육 프로그램 제공, 직관적인 UI/UX 설계, 다국어 지원 및 문화적 맥락을 고려한 설명, 실제 화폐와의 명확한 가치 환산 제시 등의 방법을 통해 메타버스와 가상 경제에 대한 이해도를 높일 필요가 있다.

마지막으로 현재의 기술로는 가상 아이템의 실제 가치를 정확히 측정하고 실시간으로 반영하는 데 한계가 있을 수 있다. 가상 경제의 시장 변동성, 대규모 데이터의 실시간 처

리의 어려움, 플랫폼 간 기준차이로 인한 범용적 가치 측정 기준 마련의 어려움 등의 기술적 난관이 실제 가치를 산출하는 데 장애가 될 수 있다. 향후에 블록체인 기술을 활용한 투명하고 신뢰할 수 있는 거래 기록 시스템 구축, AI 기술을 활용한 실시간 가치 평가 알고리즘 개발 메타버스 플랫폼 간 표준화된 가치 평가 시스템 구축, 규제 기관과의 협력을 통한 가상 자산 가치 평가 기준 마련 등의 노력을 통해 가상 아이템의 가치를 더욱 정확하게 측정할 수 있는 기술 발전과 제도적 정비를 이끌어 낼 필요가 있을 것이다.

마. 데이터통제권 보장

이용자 개인정보 관리 부문에서는 이용자가 직접 제공하는 정보(5/5), 서비스 사용 시 자동으로 수집되는 정보(5/5), 이용자 개인정보 처리 목적(5/5)에 대해 매우 명확하게 안내하고 있다고 응답하였다. 또한 개인정보 이용자 권리 행사 및 문의 시 연락 방법에 대한 안내(5/5)도 긍정적으로 평가되었다. 이는 사업자 B이 이용자의 개인정보 관련 문의나 요청을 처리하는 채널은 잘 구축하고 있고 개인정보 수집과 이용에 대한 투명성을 높이는 노력이 중요하다고 인식하고 있음을 보여준다.

〈표 5-14〉 데이터통제권 보장 관련 조사 내용－사업자 B

번호	항목	결과
E1	이용자가 직접 제공하는 정보에 대한 명확한 안내	매우 그렇다(5/5)
E2	서비스 사용 시 자동으로 수집되는 정보에 대한 명확한 안내	매우 그렇다(5/5)
E3	이용자 개인정보 처리 목적에 대한 명확한 안내	매우 그렇다(5/5)
E4	이용자의 개인정보 열람 및 수정 방법에 대한 쉽고 명확한 안내	그렇지 않은 편이다(2/5)
E5	개인정보 관련 이용자 권리에 대한 명확한 안내	그렇지 않은 편이다(2/5)
E5-1	개인정보 이용자 권리 행사 및 문의 시 연락 방법에 대한 쉽고 명확한 안내	매우 그렇다(5/5)

반면 일부 영역에서는 개선이 필요한 것으로 나타났다. 이용자가 자신의 개인정보를 열람하고 수정할 수 있는 방법에 대한 안내(2/5)와 개인정보 관련 이용자 권리에 대한 안내(2/5)는 상대적으로 낮게 평가했다. 이는 사업자가 개인정보 열람 및 수정 방법 및 개인정

보 관련 이용자 권리 안내 원칙에 대한 필요성을 상대적으로 낮게 인식했기 때문일 수 있다. 대부분의 이용자들은 개인정보 열람 및 수정 과정을 경험하지 않기 때문에 관심이 낮고 필요성을 느끼지 못한다. 이러한 상황에서 개인정보 관련 안내를 상세히 제공하면 오히려 사용자 경험이 복잡해질 수 있고 이용자의 피로를 유발할 수 있다는 우려 때문에 이를 최소화하려 했을 수 있다. 또한 메타버스 환경에서는 아바타, 가상 자산 등 다양한 형태의 개인정보가 존재한다. 이러한 복잡성으로 인해 일괄적인 개인정보 열람 및 수정 방법을 제시하기 어려웠을 수 있다.

그럼에도 불구하고 개인정보보호법이 규정하는 바와 같이 정보주체의 권리를 명확히 하고 이에 대한 안내는 이용자 권익 보호를 위한 필수적인 조치이므로 이용자 사용경험을 방해하지 않는 선에서 안내 의무를 준수할 필요가 있을 것이다. 사업자는 이용자의 개인정보 관련 권리가 보호될 때의 명확한 이점과 보호되지 않을 때 나타나는 문제들을 구체적으로 제시하여 이러한 규정에 대해 이용자들의 관심을 환기시킬 필요도 있다.

바. 지속가능발전 지향

지속가능발전 지향에 대한 사업자 B의 자체 진단 설문 결과를 보면, 메타버스 서비스의 지속가능성과 이용자에 대한 장기적 영향을 고려하고 있지만, 아직 체계적이고 지속적인 평가와 대응 시스템을 구축하지 못하고 있음이 나타났다. 구체적으로, 메타버스 서비스가 이용자에게 미치는 영향을 평가하고 대응하기 위한 노력에 대해서는 ‘보통이다’(3/5)라고 응답하면서 게시물 피드 기능을 운영하는 과정에서 쾌적한 이용 환경을 만들기 위해 불쾌감이나 혐오감을 주는 게시물을 자동으로 필터링하고, 운영진이 모든 게시물에 대해 2차 검수를 실시하여 추가적인 관리를 하고 있다는 점을 구체적인 노력의 사례로 제시하였다.

〈표 5-15〉 지속가능발전 지향 관련 조사 내용 - 사업자 B

번호	항목	결과
F1	메타버스 서비스가 이용자에게 미치는 영향을 평가하고 대응하기 위한 노력	보통이다(3/5)
F1-1	구체적인 노력 사례	게시물 피드 기능을 운영하는 과정에서 쾌적한 이용 환경을 만들기 위해, 불쾌감/혐오감을 주는 게시물을 자동 필터링하고 있으며, 이후 운영진이 전수 검사를 통해 모든 게시물에 대해 2차 검수를 하고 있음
F1-2	이용자에게 미치는 영향에 대한 평가와 대응의 정례적 실시	그렇지 않은 편이다(2/5)
F1-3	이용자에게 미치는 영향에 대한 평가와 대응의 내부구성원 공개	그렇지 않은 편이다(2/5)
F1-4	이용자에게 미치는 영향에 대한 평가와 대응에 대한 이용자 피드백 수집	그렇지 않은 편이다(2/5)

그러나 이용자에게 미치는 영향에 대한 평가와 대응이 정례적으로 이루어지지 않고 있다는 점(2/5), 평가와 대응 결과가 내부 구성원에게 충분히 공개되지 않는다는 점(2/5), 이용자에게 미치는 영향에 대한 평가와 대응에 대해 이용자의 피드백을 받는 것이 미흡한 수준이라는 점(2/5) 등에서 부정적으로 평가했으며 개선이 필요한 사항으로 나타났다.

이러한 결과는 메타버스 서비스가 이용자에게 미치는 영향을 평가하고 대응하기 위한 노력이 전반적으로 매우 부족함을 보여주었다. 이는 메타버스 산업이 아직 초기 단계에 있어 서비스 개발과 사용자 유치에 주력하고 있음을 감안하더라도 메타버스 사업자들이 서비스의 장기적 영향과 잠재적 위험에는 어떤 것들이 있으며 이에 어떻게 대응해야 하는지가 중요하다는 사실을 인식하지 못하고 있고, 이를 체계적으로 연구하고 평가하려는 노력도 부족함을 시사한다.

메타버스 서비스가 이용자에게 미치는 영향을 평가하고 대응하는 노력은 매우 중요하다. 메타버스는 현실과 가상의 경계를 넘나드는 새로운 경험을 제공하므로, 이용자의 심리적, 사회적, 경제적 안전을 보장하기 위해 지금까지 알려지지 않은 잠재된 문제들 및 역기

능을 파악하고 선제적으로 대응하는 노력이 필요하다는 것이다. 이는 메타버스 이용 과정에서 나타날 수 있는 장·단기적 효과 연구 및 부정적 효과 예방을 위한 정책 연구 등 연구지원 강화를 통해 가능하다. 메타버스가 사회적으로 유익한 서비스를 제공하고 책임 있게 운영하기 위해서는 이러한 투자와 노력을 통해 사회로부터 신뢰를 얻어야 한다.

구체적으로 분기별 또는 반기별로 이용자 영향을 평가하고 결과를 분석하는 정기적이고 체계적인 연구 시스템을 마련해야 한다. 연구 시스템 안에는 내부 구성원뿐만 아니라 이용자, 전문가, 시민단체 등 다양한 이해관계자가 참여해서 평가하고 대응 프로세스를 구축하고, 평가 결과와 대응 방안을 내외 관계자들 및 이용자들에게 투명하게 공개하며, 피드백을 적극적으로 수렴해야 한다. 메타버스 업계가 협력하여 이용자 영향 평가 및 대응에 대한 공통 가이드라인을 마련하고 이를 잘 준수하는 모범 사례를 개발하는 것도 중요하다. 정부는 메타버스 사업자들의 이용자 영향 평가 및 대응 노력을 지원하고, 필요한 경우 관련 규제를 마련해야 한다. 메타버스 서비스 제공자와 이용자 모두에게 잠재적 위험과 영향에 대한 교육도 뒷받침되어야 메타버스 서비스의 지속가능한 발전과 이용자 보호를 동시에 달성할 수 있을 것이다.

사. 정부역할 - 서비스 제공자 지원 및 이용자 역량 강화와 보호

사업자의 서비스 제공자 지원 및 이용자 역량 강화와 보호 측면에서 정부역할에 대한 의견은 다음과 같았다. 먼저 서비스 제공자 지원을 위한 정부의 적극적인 역할이 필요하다고 보았다. 구체적으로, 서비스 제공자에 대한 정부의 지원 정책 수립(5/5)과 자율규제 활동을 위한 정부의 구체적인 실행 가이드 제공의 필요성에 매우 동의했다(5/5). 기본원칙에 대한 실행 가이드 제공(4/5), 기본원칙 준수 유무 관리 감독(4/5), 정부가 서비스 이용자의 인공지능 이용 역량 강화를 위한 지원을 해야 한다는 점(4/5), 서비스 제공자의 이용자 보호 책무 이행을 촉구해야 한다는 데(4/5)에 대해서도 동의한다는 입장을 보였다.

〈표 5-16〉 정부역할 관련 조사 내용－사업자 B

번호	항목	결과
G1-1	서비스 제공자에게 기본원칙에 대한 실행 가이드 제공	약간 동의한다(4/5)
G1-2	서비스 제공자에 대한 지원 정책 수립	매우 동의한다(5/5)
G1-3	서비스 제공자의 자율규제 활동을 위한 구체적인 실행 가이드 제공	매우 동의한다(5/5)
G1-4	서비스 제공자의 기본원칙 준수 유무 관리 감독	약간 동의한다(4/5)
G1-5	정부역할 관련 추가 의견	여러 메타버스 서비스를 포괄할 수 있는 표준 가이드를 제공하되, 서비스 별로 특성과 상황이 다양하므로 너무 획일적이거나 엄격한 규제, 감독은 지양되면 좋겠습니다.
G2-1	서비스 이용자의 인공지능 이용 역량 강화를 위한 지원	약간 동의한다(4/5)
G2-2	서비스 제공자의 이용자 보호 책무 이행 촉구	약간 동의한다(4/5)
G2-3	정부역할 관련 추가 의견	인공지능 이용 역량이 중요한 시대입니다. 다만 이용자의 역량과 관련해서는 이용자 자율에 맡겨야 합니다.

전반적으로 서비스 제공자를 지원하고 관리 감독하는 정부의 역할을 중요하게 생각하지만, 서비스별 특성과 상황이 다양하므로 너무 획일적이거나 엄격한 규제, 감독은 지양해야 한다고 의견을 제시했다. 또한 이용자 역량 강화 및 보호를 위한 정부 지원의 필요성을 인정하면서도 인공지능 시대에 맞게 이용자의 역량개발은 각자의 자율에 맡기는 것이 보다 현실적이라는 의견을 제시하였다. 또한 여러 메타버스 서비스를 포괄할 수 있는 표준 가이드를 제공해줄 것을 주문하기도 했다.

제5절 시사점

1. 설문조사 결과에 대한 논의

가. 긍정적인 항목

메타버스 사업자 A와 사업자 B는 모두 다양한 유해콘텐츠를 적극적으로 제한하면서 ‘공동체 가치 형성’에 대한 기본원칙을 철저하게 지키고 있다고 응답하였다. 이용자 콘텐츠에 대한 개념, 권리, 활용, 분쟁 조정 방식 등 콘텐츠와 관련한 이용자 권리에 대해서도 명확하게 안내하고 있으며, 이용자들의 궁금증이나 불만사항을 접수/해결하고, 이용자들 간 갈등과 분쟁을 조정하는 체계도 잘 갖추고 있다고 자체 평가하였다. 두 사업자들은 이용자 개인의 활동정보의 수집과 활용에 대해서도 자세히 안내하고 있다고 응답했다. 마지막으로, 두 사업자 모두 이용자 보호를 위한 정부의 적극적인 역할이 필요하다는 입장을 보였다.

나. 개선이 필요한 항목

1) 2차 생산물 제작 및 활용에 대한 안내

사용자 콘텐츠에서 비롯된 2차 생산물에 대한 안내는 두 사업자 모두 전혀 이루어지지 않고 있었다. 메타버스에서 ‘사용자 콘텐츠를 통한 2차 생산물’은 사용자가 직접 만든 콘텐츠를 기반으로 새롭게 창작되거나 변형된 콘텐츠를 의미한다. 다른 사용자가 만든 아바타 의상이나 액세서리를 구매하여 자신의 아바타에 적용하는 아바타 커스터마이징, 다른 사용자가 만든 가상 건물이나 환경을 구매하여 자신의 가상공간에 배치하거나 수정하는 가상공간 제작, 다른 사용자가 만든 게임 맵을 수정하거나 확장하여 새로운 게임을 만드는 게임 맵 제작, 다른 사용자가 만든 디지털 아트를 구매하여 자신의 가상 갤러리에 전시하거나 수정하는 디지털 아트워크 등이 2차 생산물의 예이다.

메타버스에서 사용자들이 직접 만든 콘텐츠를 기반으로 새로운 가치를 창출하는 2차 생산물은 메타버스 내 경제의 성장과 발전을 위해 중요한 동력이 될 수 있다. 즉, 다른 사용자의 콘텐츠를 활용하여 새로운 것을 만들 수 있다는 사실은 사용자들의 창작 의욕을 고취시키고, 플랫폼 내 콘텐츠의 다양성과 품질을 향상시킬 수 있으며, 콘텐츠 제작자들은 수익을 창출할 수 있기 때문에 메타버스 생태계를 발전적으로 순환시킬 수 있는 원동력이 된다는 것이다. 따라서 메타버스 사업자들은 2차 생산물 시장에 관심을 기울이고 명확한

정책과 가이드라인을 수립하여 사용자들에게 적극적으로 안내할 필요가 있다.

2) 가상 콘텐츠 및 아이템 거래에 대한 안내

사업자 A는 가상 콘텐츠 및 아이템 거래에 대한 안내를 아직 하고 있지 않은 것으로 나타났다. 가상 콘텐츠와 아이템의 거래가 실제 돈을 포함한 경제 활동과 연결되어 있음에도 불구하고, 가상 거래의 정의, 가치, 허용 범위, 저작권 및 상표권 침해 등에 대한 명확한 안내가 없다면 이용자들이 거래되는 아이템의 가치나 신뢰도를 판단하기 어렵게 되어 비정상적인 가격 책정, 거래 사기, 가치 하락 등 경제적 피해로 이어질 수 있다. 또한 가상 콘텐츠 및 아이템 거래의 범위(예: 재판매 가능 여부, 수정 가능 여부)가 명확히 안내되지 않으면 사용자들이 적절한 거래와 부적절한 거래를 구분하기 어려워 사용자가 예상하지 못한 규정 위반으로 콘텐츠를 압류당하거나 계정 정지를 당하는 등 불이익을 겪을 수도 있다. 가상 거래와 관련된 법적 책임(예: 세금, 개인정보 보호, 소비자 보호)이 명확히 정의되지 않으면, 이용자와 사업자 모두 법적 문제에 휘말릴 가능성도 있다. 사용자들이 가상 콘텐츠를 제작하거나 거래하는 과정에서 저작권 또는 상표권을 침해할 가능성도 존재하는데, 이와 관련된 안내가 없다면 사용자들은 어떤 행동이 불법인지 알지 못한 채 법적 문제에 직면할 위험도 있다.

따라서 사업자는 가상 콘텐츠 및 아이템의 거래 정의, 가치 평가 기준, 허용 범위 등을 명확히 문서화하여 사용자들에게 제공하고, 저작권 및 상표권 관련 교육 및 관련 안내를 제공하여 저작권과 상표권 침해를 방지할 필요가 있다. 또한 사용자 간 거래를 보증하고, 안전한 결제를 위한 중개 서비스 및 환불 정책을 도입하고, 세금, 개인정보 보호, 소비자 보호 등 가상 거래와 관련된 법적 요구사항을 준수하면서 이를 사용자에게 명확히 공지해야 한다. 거래와 관련된 분쟁이 발생할 경우 이를 공정하고 효율적으로 해결할 수 있는 분쟁 해결 프로세스를 마련하는 일도 중요하다.

3) 이용자 활동정보에 대한 권리 안내

사업자 B는 이용자가 메타버스에서 자신의 활동내역이 담긴 개인정보를 열람하고 수정하는 방법을 명확하게 안내하고 있지 않고, 개인정보 이용자 권리 행사 및 문의를 위해서 연락 방법을 자세히 안내하고 있지 않다고 응답했다. 사용자는 자신의 활동 내역과 관련된 개인정보를 열람하거나 수정할 권리를 가지고 있으며, 이를 쉽게 행사할 수 없으면 신

위가 저하되고 플랫폼 이용을 꺼리게 될 가능성이 높다. 개인정보보호법에서는 이용자에게 개인정보 접근 및 수정 권한을 명확히 제공하도록 요구하고 있는데, 이를 제대로 안내하지 않으면 법적 책임을 초래할 수 있다. 무엇보다 개인정보 수정 및 문의 방법이 명확히 제공되지 않으면 사용자는 불편함을 느끼며, 플랫폼에 대한 신뢰가 약화될 수 있다.

따라서 사업자들은 이용자가 자신의 개인정보를 쉽게 열람하고 수정할 수 있도록, 플랫폼 내에 직관적인 인터페이스와 상세한 안내문을 제공해야 한다. 개인정보 처리방침에 열람, 수정, 삭제 절차를 구체적으로 명시하고, 사용자가 이를 쉽게 이해할 수 있도록 FAQ나 가이드 문서를 제공하고, 온라인 고객센터와 문의 채널을 연결하여 즉각적인 지원이 가능하도록 해야 한다. 또한 이메일, 전화, 실시간 채팅 등 다양한 방법을 제공하고, 응답 시간에 대한 SLA(Service Level Agreement)를 설정하여 신속히 대응할 필요가 있다.

4) 지속가능한 발전을 위한 노력

메타버스 서비스가 이용자에게 미치는 영향을 평가하고 대응하는 것은 매우 중요하며, 이를 위해 정기적이고 체계적인 연구 시스템 구축이 필요하다. 분기별 또는 반기별로 이용자 영향을 평가하고, 평가 결과를 내외부 관계자들과 투명하게 공유하며, 피드백을 반영하는 프로세스를 마련하는 것은 이용자 신뢰를 구축하는 데 핵심적인 역할을 할 것이다. 특히, 연구 시스템에는 이용자, 전문가, 시민단체 등 다양한 이해관계자가 참여해 객관성을 확보하고, 메타버스 서비스가 미칠 수 있는 단기적·장기적 효과를 다각도로 분석할 필요가 있다. 더불어 메타버스 업계가 협력하여 이용자 영향 평가 및 대응에 대한 공통 가이드라인을 마련하고 모범 사례를 제시하는 것은 플랫폼의 책임성을 강화하는 중요한 방안이다. 이러한 노력은 정부의 지원 및 필요한 규제와 함께 이루어질 때 더욱 효과적일 것이다.

5) 이용자 보호를 위한 정부의 역할

정부는 서비스 별로 특성과 상황이 다양하다는 점을 고려하여 여러 메타버스 서비스를 포괄할 수 있는 표준 가이드와 기본원칙에 대한 구체적인 실행 가이드를 사업자들에게 제공하고 관리 감독하며 필요한 지원정책을 수립해야 한다. 또한 취약계층을 포함하여 이용자의 AI 리터러시 역량을 개발하는 데 사업자와 협력해야 한다.

사업자들이 이러한 기본원칙과 실행 가이드를 효과적으로 이행하고 있는지 그 현황을

데이터 기반으로 정기적으로 점검하고 평가하는 시스템을 마련하고, 우수 사업자에게는 인증을 부여해 긍정적 경쟁을 유도하는 방안을 마련해야 한다. 또한, 이용자, 전문가, 시민 단체가 사업자의 실행여부를 평가할 수 있는 공개 플랫폼을 제공하고, 독립적인 감독 기구를 설립하여 공정성과 투명성을 확보하는 것도 필요할 것으로 보인다.

정부는 이용자 보호 기술 개발과 취약계층 지원 프로그램 운영을 위한 재정적 지원과 세제 혜택을 제공해야 하며, 공통 기술 표준과 가이드라인을 마련해 사업자들이 이를 효과적으로 적용할 수 있도록 해야 한다. 또한, 사업자 간 협력을 지원해 유해 콘텐츠 관리와 개인정보 보호 등 공통 과제를 함께 해결할 수 있는 환경을 조성해야 한다. 이러한 조치를 통해 정부와 메타버스 사업자는 이용자 보호와 서비스 혁신을 동시에 이루며 지속 가능한 메타버스 생태계를 구축할 수 있을 것이다.

2. 인터뷰 결과에 대한 논의

인터뷰 과정에서 알고리즘 추천 서비스 이용자 보호를 위한 기본원칙과 관련하여 논의된 몇 가지 쟁점들을 정리하면 다음과 같다.

첫째, 서비스 추천 알고리즘에 대한 공개의 범위를 지나치게 상세한 수준으로 권고하는 것이 적절하지 않다는 의견이 있었다. AI 기술이 고도화됨에 따라 알고리즘이 복잡하고, 더욱이 AI 학습의 결과로 생성된 추천서비스를 구성하는 변수와 파라미터는 소위 ‘블랙박스’ 안에 들어 있는 요소로 이를 구체적으로 파악할 수 없기 때문에 명확한 기준을 설명할 수도 없다는 것이다. 또한 AI 기술발전 속도가 빠르기 때문에 이를 적용하여 새로운 추천 서비스가 개발될 때마다 이용자들에게 공지하기보다 수집항목, 이용목적, 보관 기간 등 주요 변경사항이 있을 때 통지하는 것을 원칙으로 삼는 것이 적절할 것으로 판단된다.

둘째, 개발부서 별로 가지고 있는 내부지침에 대한 공개 요구도 재고할 필요가 있다. 사업자 입장에서 내부지침은 변동 상황이 자주 발생하고 고유의 사업전략이 포함되기도 하므로 왜 공개해야 하는지 이해가 부족한 상황인데, 만약 공개 방침이 계속된다면 해외 사업자들은 이를 이행하지 않을 가능성이 높으므로 국내 사업자들에게 불이익이 갈 수 있다. 따라서 내부지침에 대한 공개가 필요하다면 이에 대한 명확한 이유, 범위, 내용이 제시되어야 할 것으로 보인다.

셋째, 현재 사이트 내에서의 이용자 활동정보를 수집하여 이를 기반으로 개인화 된 추

천 서비스를 제공하고 있는데, 수집되는 이용자 활동정보와 추천 서비스의 세부사항을 이용자들이 쉽게 선택할 수 있는 방법이 제공될 필요가 있다고 판단한다. 현재는 로그인을 하게 되면 자동적으로 개인활동 정보가 수집되고 개인 맞춤형 추천 서비스가 제공되는데, 이용자들이 세부사항을 선택할 수 있도록 이러한 기능을 가진 메뉴를 로그인 후 ID 정보 옆에 쉽게 찾을 수 있는 위치에 제공할 필요가 있다.

넷째, 알고리즘 추천 서비스 사업자가 해결해야 할 또 하나의 중요한 문제 중 하나는 유해표현 걸러내는 시스템을 개발하는 것인데, 현재는 단어 기반 제어 장치가 작동되고 있으나 이를 넘어 맥락까지 잡아낼 수 있는 시스템 개발이 필요하다. 최근 고성능 언어모델이 개발되어 기계학습의 정확도가 획기적으로 개선됨에 따라 정확도 높은 분류기 개발이 가능해졌다. 유해표현에 대한 명확한 개념정의와 함께 범주, 대상, 강도/심각성 등을 기준으로 명확한 분류체계를 만들고 정확도 높은 훈련데이터를 구축하는 연구를 적극적으로 지원할 필요가 있다.

마지막으로 취약계층을 포함한 이용자의 AI 역량 강화를 위해 더 많은 노력을 기울일 필요가 있다. 정보와 기술 격차는 심각한 문제로, AI와 같은 고도화된 기술은 이러한 격차를 더욱 심화시킬 수 있으며, AI 활용 능력은 이제 모든 사람들이 보편적으로 갖춰야 할 필수 역량이 되었다. 특히, 고령자나 저소득층 등 취약계층이 디지털 환경에서 소외되지 않도록 AI 관련 교육과 지원이 절실한 상황이다.

사업자는 AI 교육 및 접근성 강화를 통해 사회적 문제 해결에 기여할 책임이 있다. 그러나 치열한 시장 경쟁 속에서 사업자들은 사회적 역할보다 성과 중심의 전략에 집중하는 경향이 있다. 이는 단기적으로는 효율적일 수 있지만, 장기적으로는 사회적 신뢰와 기업의 지속 가능성을 저하시킬 위험이 있다. 사업자들은 AI 역량 강화를 위해 간단하고 직관적이면서 일상생활에서 직접적으로 도움이 되는 AI 활용 방법을 중심으로 교육하고, 공공기관과 협력해 확대하고 정부 지원 프로그램과 연계하여 무료 교육 세션, 디지털 장비 제공, 지역 커뮤니티 중심의 교육 허브를 운영하는 등의 노력을 기울일 필요가 있다. 여러 기업이 공동으로 AI 교육 플랫폼을 개발하거나, 공유 가능한 교육 콘텐츠 제작하는 것도 좋은 방법일 것이다. AI 교육 지원을 기업의 사회적 책임(CSR)과 연계하여 브랜드 신뢰를 구축하고 장기적 성장을 도모, AI 교육에 투자함으로써 디지털 환경에서의 포용성을 확대하고, 이는 궁극적으로 더 넓은 사용자층을 확보하는 데 기여할 수 있을 것이다.

제6장 결 론

본 보고서는 2019년 방송통신위원회에서 발표한 「이용자 중심의 지능정보사회를 위한 원칙」 이후 후속적으로 추진된 정책개발 과정과 성과를 담았다. 먼저 2장부터 3장까지는 신규 서비스로 새로운 사회 이슈를 발생시키고, 사람들의 소통 방식이나 정보 검색, 콘텐츠 소비 방식에 큰 변화를 가져온 생성형 인공지능 서비스에 대한 이용자 보호 가이드라인을 개발하는 과정과 성과를 다루었다. 인공지능은 인간의 기본권을 침해할 수 있는 잠재적 위험을 가진 기술이다. 인공지능 기술 특성상 데이터 학습 과정을 거쳐야만 하는데, 이 학습을 위한 데이터가 인간 스스로 기본권을 확립해 가는 동안 발생시킨 온갖 오류와 편향을 담은 데이터이기 때문이다. 따라서 어떤 인공지능 모델이든, 데이터 학습 과정을 거친다면 인간의 기본권 침해를 최소화하는 방안을 마련해야 한다. 본 가이드라인도 이 인격권 보호를 최우선 지침으로 제시하고 있다. 또 생성형 인공지능 서비스를 이용할 때, 해당 콘텐츠가 인공지능에 의해 생성된 콘텐츠라는 인지와 판단이 중요하다. 이러한 인지와 판단을 통해, 인간 고유의 권리인 저작권의 가치를 판단할 수 있고, 또 실재하지 않은 허위 콘텐츠 여부도 판단할 수 있다. 또 자연스럽게 상호작용하는 대상이 인공지능이라고 인지하는 것 역시 과도한 기술 도구 사용에 따른 중독 현상을 경계하고 객관적 판단력을 흐리는 우를 방지할 수 있다고 본다. 이러한 차원에서 투명성과 설명가능성을 두 번째 주요 지침으로 제시하였다.

세 번째는 인간사회가 긴 역사를 거치면서 정제해 오고 있었던 편견과 차별이 인공지능 생성물에서 반복되거나 강화되는 오류가 발생하지 않도록 점검하고 사전, 사후 모니터링과 필터링 체계 등을 강조하는 다양성 존중에 대한 지침을 제시하였다. 네 번째는 이용자들이 생성형 인공지능과 상호작용하면서 발생하고 축적하는 데이터를 다시 학습데이터로 활용하였을 때 발생할 수 있는 잠재적 위험을 사전에 차단하기 위한 지침에 대한 것이다. 생성형 인공지능 서비스를 구현하는 핵심 모델은 사전에 대규모 언어 학습을 하는 LLM(거대언어모델)이지만, 이 이슈에서 자유롭지 않았다. 해당 인공지능 모델 등장 이후 2년이 채 안 되어 해당 모델의 약점을 보완하기 위한 방안으로 파인튜닝 및 다른 학습 및 테

이더 처리 과정이 결합되는 등의 서비스 진화가 일어났다. 이 과정에서 이용자의 서비스 목적에 따라 한정된 기기와 시기 동안 이용자와의 상호작용 동안 생성된 이용자 데이터를 재학습하여 개인화 서비스를 제공하는 서비스 방식이 있었다. 현 시점에서는 그 순환 과정이 복잡하지 않지만 향후 순환의 기간이 길어지고 데이터 축적과 활용이 반복될수록 이용자의 지극히 사적인 데이터가 온라인 상에 존재하게 되거나 제한된 범위 내 반복된 학습으로 왜곡된 결과를 초래하는 등의 부가 이슈가 발생할 수 있다는 점을 염두에 둘 필요가 있었다.

다섯 번째와 여섯 번째 지침은 이용자가 생성형 인공지능 서비스를 어떻게 활용하느냐에 따라 잠재적 위험이 발현될 수도 혹은 그렇지 않을 수도 있는 이슈들을 다룬다. 이는 본 가이드라인 개발 전에 발간되었던 기존 생성형 인공지능 관련 가이드라인 대부분이 이용자에게 지침을 주는 활용 가이드였던 이유이기도 하다. 이용자의 활용 지침에 대한 가이드는 기계발된 가이드라인에서 습득할 것을 전제로 하고, 본 가이드라인에서는 사업자가 선제적으로 이용자의 오남용 가능성을 사전에 최소화할 수 있는 요소가 있는지를 살폈고, 그 요소들을 기준으로 하였다. 그 요소들에는 생성 콘텐츠 활용에서 발생할 수 있는 문제에 책임 있게 대응하는 태도, 그리고 생성 콘텐츠의 건전한 유통과 배포를 위한 노력이 꼽혔다.

다섯 번째 지침으로 활용에서 발생할 수 있는 문제에 대한 책임과 참여 주제에서는 주요 행위자의 책임 범위를 분명히 할 필요가 있음을 권고하고, 상대적으로 정보력이 낮은 이용자의 오남용을 관리하기 위한 모니터링 시스템 등의 위험관리 체계 마련을 요청한다. 여섯 번째 지침은 이용자가 생성하거나 생성하려는 콘텐츠가 도덕적, 윤리적 기준에 부합한지를 판단하고 그에 따른 대응을 하도록 권고한다. 이러한 지침은 현재 널리 활용되고 있는 인공지능 윤리기준에 준하거나 청소년보호법의 청소년 유해매체물 제공·판매·배포 등의 기준에 준하여 기술하였다.

모든 변수를 다 고려한 완벽한 가이드라인을 개발하는 데는 한계가 있을 수 있으나, 사회적으로 이슈가 된 문제들을 두루 다루고 사업자들이 실천할 수 있는 합리적 수준에서 최소한의 기준을 마련하고자 하였다. 이를 위해, 컴퓨터 공학자, 법조계 및 법학 전문가, 소비자학 및 미디어학 전문가 등 다양한 분야 전문가들과 논의하고, 사업자들의 의견을 듣는 등 전문가 및 이해당사자의 의견을 수렴하는 절차를 거쳤다. 이 과정에서 몇 가지 기

준들은 완화되거나 빠지기도 하였다.

예를 들어, 투명성 지침의 경우, 아직 명확하게 어떤 정보를 꼭 포함해야 한다는 법적 기준(강제력이 있는 기준)이 부재하여 사업자의 자율 판단에 맡길 수 밖에 없었으며, 인공지능이 생성한 콘텐츠임을 표시하는 보편적인 방법인 디지털 워터마킹에 대한 한계를 고려하여 완화된 지침을 제시할 수밖에 없었다. 사업자들에 따르면, 현재 디지털 워터마킹 기술은 콘텐츠 원본이나 서비스의 질을 떨어뜨리지 않고 효용성 있게 활용할 수준에 도달하지 못하였다. 이에 본 가이드라인에서는 현재의 수준을 정보 형태로 알리고 사업자의 노력과 향후 기술 및 제도 발전에 따라 기준이 높아질 필요가 있음을 제시하는 정도로 하고 현재 수준에서 실천가능한 수준으로 지침을 구성하였다.

인공지능과 상호작용하면서 생성한 개인의 데이터를 학습데이터로 활용하는 이슈에 대해서도, 현재 시점에서는 해당 서비스 상황이 활성화된 단계가 아니라서 지침으로 넣는 것에 대한 이의가 있었다. 그러나 일부 글로벌 서비스가 해당 부문에 대한 이용자 정책을 실행하고 있어서 빠지 않고 조건을 붙여 담았다. 즉, '서비스 이용 과정에서 생성된 고객 데이터를 활용하는 경우'와 같은 조건을 붙인 것이다.

이번에 개발한 「생성형 인공지능 서비스 이용자 보호 가이드라인(안)」은 일정 주기로 추적, 보완해야 할 자율규범(안)으로 이해할 필요가 있다. 기본적으로 계속 변화 및 발전하는 생성형 인공지능 서비스를 대상으로 가이드라인을 개발하면서 변화 및 발전 중에도 변치 않고 꼭 지켜야 할 최소한의 필수 기준을 지침으로 담으려고 하였고, 다른 한편에서는 현재 시점에서는 불필요할 수 있는 잠재적 위험에 대해서는 미래 어느 시점에 수정, 보완하는 것을 전제로 마무리 지었기 때문이다.

한편, 4장과 5장은 인공지능 기반 미디어 추천 서비스 이용자 보호 기본원칙」(19) 및 「메타버스 이용자 보호 기본원칙」(23) 수립 이후 기업 적용 현황 파악 및 인공지능 기술 발전과 이용 환경 변화에 따른 수정 및 보완 사항을 진단한 연구 과정과 결과를 다루었다. 4장에서는 주요 미디어 플랫폼의 인공지능 기반 추천 서비스를 대상으로 투명성, 공정성, 책무성이라는 핵심 원칙과 함께 정보 공개, 선택권 보장, 자율검증 실행, 불만 처리 및 분쟁 해결, 내부 규칙 제정이라는 실행 원칙의 이행 수준을 체계화된 기준으로 진단하였다. 이를 위해 설문조사와 인터뷰를 실시하고 서비스 약관 및 고객센터 웹페이지를 분석함으로써 사업자의 실제 운영 방식을 입체적으로 파악했다. 이러한 종합적 분석을 통해 플랫

품별 실천 단계의 차이점을 확인하고 개선이 필요한 영역을 확인할 수 있었으며, 주요 플랫폼의 선도 사례와 모범 운영 방식도 발굴할 수 있었다. 특히 서비스 약관과 고객센터 웹 페이지에 명시된 이용자 불만 처리 절차와 추천 알고리즘의 투명성 수준을 점검함으로써, 사업자가 외부에 공시한 정책과 실제 서비스 운영이 얼마나 일치하는지 파악할 수 있었다. 본 연구의 분석 결과는 향후 관련 정책 수립을 위한 중요한 근거자료로 활용될 수 있을 뿐만 아니라, 사업자들의 자율적인 서비스 개선을 유도하는 데에도 도움이 될 것으로 기대한다.

유사한 맥락에서 5장에서는 메타버스의 규제 이행 실태를 종합적으로 점검하여 이용자 보호 기본원칙의 실효성을 평가하고 정책적 시사점을 도출하였다. 이에 본 연구는 메타버스 플랫폼 서비스 제공자들의 이용자 보호 기본원칙 이행 현황을 진단하고, 자율규제의 성과와 제약요인을 분석하였다. 이를 통해 이용자 권익 보호와 디지털 생태계의 신뢰도 향상을 정책 방향을 도출하였다. 또 메타버스 플랫폼의 사회적 영향력과 책무를 재조명함으로써, 디지털 기술이 공동체의 가치 형성과 유지에 이바지할 수 있는 발전 방향을 제시하였다. 이는 단순한 기술적 효율성을 넘어 인간 중심의 디지털 환경 구축과 지속 가능한 디지털 생태계 조성에 기여할 것으로 예상된다. 본 연구의 성과는 향후 학계와 업계, 정책 결정자들이 메타버스의 바람직한 발전 방향을 모색하는 데 있어 실질적 함의를 제공할 것으로 전망된다.

참 고 문 헌

[국내 문헌]

- 국가사이버보안센터(2023, 6). 『챗GPT 등 생성형 AI 활용 보안 가이드라인』.
- 김수정(2022). “한국의 포털 사이트 추천 알고리즘의 공정성 평가”. 《정보과학회지》, 30(2), 123-135. <<https://doi.org/10.5574/KJCS.2022.30.2.123>>.
- 김정수(2023). “의료에서 메타버스 활용과 법적 쟁점”, 《저스티스》, (194-3), 233-263.
- 김정환(2023). “메타버스와 온라인 분쟁해결 (ODR)”, 《저스티스》, (194-3), 302-334.
- 김태오(2009). “통신시장 이용자보호의 쟁점과 과제”, 《경제규제와 법》, 제2권 제1호
- 남현숙·곽나연(2023). 『주요국 메타버스 정책 동향 분석』. 소프트웨어정책연구소
<https://www.spri.kr/posts/view/23671?code=data_all>.
- 라기원(2024). 『유럽연합 인공지능법 (EU AI ACT)의 특성과 쟁점 - 우리나라 인공지능 입법에 대한 시사』, 규제혁신법제연구 이슈페이퍼, 24-20-3, 한국법제연구원.
- 박종수(2013). “ICT융합에 따른 방송통신 생성형 인공지능 서비스 이용자 보호 법제의 합리적 개선방안”, 《법제연구》, 제44호, 한국법제연구원, 108쪽 이하.
- 방송통신위원회(2021). 『인공지능 기반 미디어 추천 서비스 이용자 보호 기본원칙』
<<https://kcc.go.kr/download>>.
- _____ (2022). 『인공지능 기반 미디어 추천 서비스 이용자 보호 기본원칙 해설서』.
- _____ (2023). 『메타버스 이용자 보호 기본원칙』.
- _____. NIA(2023. 12). 『생성형 AI 윤리 가이드북』.
- 서울디지털재단(2023. 11). 『서울시 생성형AI 윤리 가이드라인』.
- 신현주(2022). “메타버스의 발전 동향 및 비즈니스 모델에 대한 연구”, 《e-비즈니스연구》, 23(1), 247-262.
- 유승현(2022). “디지털 전환 시대의 데이터 정책과 이용자 권리 보호”, 《방송통신연구》, 2022년 특집호, 38-73.

- 이상욱(2021). 『유네스코 인공지능(AI) 윤리 권고 해설서 - 인공지능 윤리 이해하기』, 유네스코한국위원회 (보고서).
- 이연경·김현지·박준우(2022). “메타버스를 이용한 버추얼 가수의 유형 제시 및 현황 연구”, 《한국엔터테인먼트산업학회논문지》, 16(6), 107-121.
- 이원태·이호영·김경훈·손상영·선지원·문정욱·김병우·황용석·최경진·심우민·한문승. (2018). 『지능정보화 이용자 기반 보호 환경 조성 총괄보고서』. 정보통신정책연구원
- 이하은·한정엽(2021). “메타버스 플랫폼의 체험형태에 따른 유형 분류 및 특성 연구”, 《한국공간디자인학회》, 16(8), 427-435.
- 이화민(2023). “생성형 AI를 넘어 AGI 시대로”, 《전기의 세계》, 제72권 제12호.
- 이효석(2021. 1. 11). “성희롱·혐오논란에 3주만에 멈춘 ‘이루다’... AI윤리 숙제 남기다” 《연합뉴스》<<https://www.yna.co.kr/view/AKR20210111155153017>>(2025. 2. 5. 최종접속).
- 이희옥(2023). “관심사 기반 플랫폼과 이용자 보호 정책에 관한 공법적 고찰”, 《공법연구》, 52(1), 145-177.
- 정보통신산업진흥원(2024). 2024년 메타버스산업 통합 사업설명회.
<<https://www.nipa.kr/home/2-1/15198>>.
- 정지은·손나연·김현정(2022). “메타버스를 활용한 문화콘텐츠 사례 연구”, 《문화산업연구》, 22(1), 201-213.
- 조연직·신소영·박현진(2022). “메타버스를 경험한 교육 분야 종사자들의 인식 및 활용 방안 탐색”, 《교육연구》, 44(1), 33-53.
- 최경진·정경오·이종관(2015). 『통신시장 이용자 보호를 위한 법제분석 - 효율적인 피해구제 방안을 중심으로 -』, 한국법제연구원.
- 한상열(2024). 『메타버스 정책 동향 및 시사점』, 한국과학기술기획평가원.
<<https://www.kistep.re.kr/gpsBoardDownload>>.
- 행정안전부(2023. 5). 『챗GPT 활용방법 및 주의사항 안내』.
- 황용석·김기태(2019). “개인화 서비스 진전에 따른 자동추천 시스템 연구 동향과 방법론적 특성 연구”, 《사이버커뮤니케이션학보》, 36(2), 221-253.
- 황중성(2023). 『The AI Report: 美 NIST 「AI 위험관리 프레임워크(AI RMF) 1.0」 분석 및 시사점』, 한국지능정보사회진흥원.

[해외 문헌]

- Acquisti, A., Brandimarte, L., & Loewenstein, G.(2015). Privacy and human behavior in the age of information. *Science*, 347(6221), 509-514.
<<https://doi.org/10.1126/science.aaa1465>>.
- Berscheid, J., & Roewer-Despres, F.(2019). Beyond transparency. *AI Matters*, 5(2), 13 - 22.
<<https://doi.org/10.1145/3340470.3340476>>.
- Binns, R.(2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*.
<<https://doi.org/10.1145/3287560.3287583>>.
- Calvi, A., & Kotzinos, D.(2023). Enhancing AI fairness through impact assessment in the European Union: a legal and computer science perspective. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.
<<https://doi.org/10.1145/3593013.3594076>>.
- Chaudhry, M., Cukurova, M., & Luckin, R.(2022). A Transparency Index Framework for AI in Education. *ArXiv, abs/2206.03220*. <<https://doi.org/10.48550/arXiv.2206.03220>>.
- Coleman, B.(2023. 8. 18). Advancing healthcare education and clinical practice with extended reality (XR). Imperial College London News.
<<https://www.imperial.ac.uk/news/246674/advancing-healthcare-education-clinical-practice-with/>> (2025. 1. 16. 최종 접속).
- Diakopoulos, N.(2016). Accountability in algorithmic decision making. *Communications of the ACM*, 59(2), 56-62. <<https://doi.org/10.1145/2844110>>.
- Ebers, M.(2020). Regulating AI and Robotics, in: Ebers/Navas(Ed.), *Algorithms and Law*, Cambridge University Press.
- Edizel, B., Bonchi, F., Hajian, S., Panisson, A., & Tassa, T.(2019). FaiRecSys: mitigating algorithmic bias in recommender systems. *International Journal of Data Science and Analytics*, 9, 197-213.
- EU.(2024. 1. 31). Ethics guidelines for trustworthy AI | Shaping Europe's digital future.

- <<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>>.
- European Commission(2021). AI Act.
<<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>>.
- _____ (2022. October 27). The Digital Services Act.
<https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en>.
- _____ (2024, 1, 26). Proposal for 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.
<<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>>.
- _____ (2024. 7. 18.). EUROPE'S CHOICE - POLITICAL GUIDELINES FOR THE NEXT EUROPEAN COMMISSION 2024–2029 Ursula von der Leyen, Candidate for the European Commission President.
<<https://dorie.ec.europa.eu/en/details/-/card/8005305>>.
- Givens, A., & Morris, M.(2020). Centering disability perspectives in algorithmic fairness, accountability, & transparency. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. <<https://doi.org/10.1145/3351095.3375686>>.
- Jiang, Y., & Qi, Y.(2024). Design of Autonomous Learning and Communication Platform based on Interpretable Personalized Recommendation. *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, 1100–1105.
- Kaminski, M. E.(2020). Understanding Transparency in Algorithmic Accountability. In W. Barfield (Ed.), *The Cambridge Handbook of the Law of Algorithms* (pp. 121–138). chapter, Cambridge: Cambridge University Press.
- Kim, B., & Doshi-Velez, F.(2021). Machine Learning Techniques for Accountability. *AI Mag.*, 42, 47–52. <<https://doi.org/10.1002/j.2371-9621.2021.tb00010.x>>.
- Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., ... & Lomas, E.(2021). Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and

- Technological Risks of AI, ML and Associated Algorithms. *SSRN Electronic Journal*.
 <<https://doi.org/10.2139/ssrn.3778998>>.
- Noble, S. U.(2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.
- OECD.(2019. 5. 22; amended on 2024. 5. 3). Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449.
 <<https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>> (2024. 7. 22. 최종 접속).
- Pariser, E.(2011). The filter bubble: What the internet is hiding from you. Penguin Press.
- Riemer, A. R., Holland, K., McCracken, E., Dale, A., & Gervais, S. J. (2022). Does the affirmative consent standard increase the accuracy of sexual assault perceptions? It depends on how you learn about the standard. *Law and human behavior*, 46(6), 440.
- Shin, D.(2020). User perceptions of algorithmic decisions in the personalized AI system. *Journal of Broadcasting & Electronic Media*, 64(4), 541-565.
 <<https://doi.org/10.1080/08838151.2020.1843357>>.
- Sonboli, N., Smith, J., Berenfus, F., et al.(2021). Fairness and Transparency in Recommendation. *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. <<https://doi.org/10.1145/3450613.3456835>>.
- Storms, E., Alvarado, O., & Krebs, L.(2022). Transparency is Meant for Control. *Proceedings of the ACM on Human-Computer Interaction*, 6, 1-24.
 <<https://doi.org/10.1145/3555130>>.
- Susnjak, T. McIntosh, T. R., Barczak, A. L., Reyes, N. H., Liu, T., Watters, P., & Halgamuge, M. N.(2024). Over the Edge of Chaos? Excess Complexity as a Roadblock to Artificial General Intelligence. <*arXiv:2407.03652 [cs.AI]*>.
- Taylor Wessing.(2023). The EU's strategy to lead on Web 4.0 and virtual worlds.
 <<https://www.taylorwessing.com/en/insights-and-events/insights/2023/09/the-eus-strategy-to-lead-on-web-4>>.

Ursula von der Leyen.(2024,. 7, 14). Political Guidelines 2024-2029. p. 10.

<https://commission.europa.eu/document/e6cd4328-673c-4e7a-8683-f63ffb2cf648_en>.

Zarsky, T. Z.(2016). The trouble with algorithmic decisions. *Science, Technology, & Human Values*, 41(1), 118-132. <<https://doi.org/10.1177/0162243915605860>>.

Zdrok, O.(2024). Fairness Metrics in AI—Your Step-by-Step Guide to Equitable Systems. Shelf <<https://shelf.io/blog/fairness-metrics-in-ai/>>.

기타 문헌

《서울경제》(2023. 5. 17.). “티빙, 맞춤형 광고 도입 검토…수익성 확 끌어올린다”.

《시사저널e》(2020. 11. 11.). “OTT업계, 콘텐츠 추천 정확도 높이기 총력”.

《아이뉴스24》(2022. 11. 24.). “접속은 신속, 화질은 향상, 추천은 똑똑’…티빙, UX 경험 확보한다.”.

《연합뉴스》(2019). “여성, 인종 차별 조장하는 검색 알고리즘”.

<<https://www.yna.co.kr/view/AKR201908010834000005>>.

_____ (2020. 9. 28.). “토종OTT 웨이브 회원 1천 만 돌파… ‘추천 기능 최적화’”.

《중앙일보》(2023). “AI 추천 알고리즘의 투명성과 공정성 문제를 해결하기 위한 대책이 시급하다”. <<https://www.joongang.co.kr/article/23957234>>.

《테크수다》. “조성철 CTO가 전하는 티빙 고객 취향 파악하기 시즌1”.

<<https://brunch.co.kr/@techsuda/49>> (2024. 12. 23. 최종 접속).

고려대학교(2023, 3, 15). <ChatGPT 등 AI의 기본 활용 가이드라인>.

넷플릭스. 고객센터 페이지. <<https://help.netflix.com/ko>>.

_____. About Netflix. <<https://about.netflix.com/ko>>.

_____. 넷플릭스의 추천 콘텐츠 시스템 작동 방법.

<<https://help.netflix.com/legal/termsofuse>>.

_____. 개인정보 처리방침. <<https://help.netflix.com/legal/privacy>>.

디즈니플러스. 법률센터 페이지. <<https://www.disneyplus.com/legal/privacy-policy>>.

_____. 고객센터 페이지. <<https://help.disneyplus.com>>.

_____. How to Configure Disney+ Parental Controls? .

<<https://www.astrosafe.co/blog/disney-plus-parental-controls>>.

메타. 메타 뉴스룸. <<https://about.fb.com/ko/news>>.

_____. “Facebook이 어떻게 작동하는지 보다 나은 이해를 돕기 위해 서비스 약관을 업데이트 합니다.

<<https://about.fb.com/ko/news/2019/06/facebook이-어떻게-작동하는지-보다-나은-이해를-돕기-위해/>>.

_____. 커뮤니티 규정. <<https://transparency.meta.com/ko-kr/policies/community-standards/>>.

_____. AI가 Facebook과 Instagram에서 보이는 콘텐츠에 영향을 미치는 방식.

<<https://about.fb.com/ko/news/2023/06/ai가-facebook과-instagram에서-보이는-콘텐츠에-영향을-미치는-방식/>>.

_____. 메타 투명성 페이지. <<https://transparency.meta.com/>>.

_____. (2023. 10. 27). Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook. (*Updated* 2024. 06. 13).

성균관대학교(2023. 3. 10). “[Teaching Tips] 챗GPT 등 AI와 학습윤리”. 교무팀.

<<https://chatgpt.skku.edu/chatgpt/notice.do?mode=view&articleNo=149035&article.offset=30&articleLimit=10>> (2024. 12. 12. 최종 접속).

_____. 챗GPT 종합안내 홈페이지, <<https://chatgpt.skku.edu/chatgpt/index.do>>.

(2024. 12. 12. 최종 접속).

연세대학교 연구윤리센터(2024, 5). 『생성형 AI 활용 가이드라인』.

월트디즈니컴퍼니 개인정보 정책 페이지. <<https://privacy.thewaltdisneycompany.com/en>>.

웨이브. 서비스 이용약관 페이지. <<https://member.wavve.com/signup/terms>>.

_____. [공지] 웨이브 이용자 보호 비전 및 계획.

<<https://www.contentwavve.com/news/articleView.html?idxno=587>>.

_____. [웨이브톡] 이용자 취향 저격수, 데이터응용개발팀 왕성현 연구원 .

<<https://wavve-comm.tistory.com/139>>.

_____. 사이버 감사실. <<https://www.contentwavve.com/com/com-5.html>>.

중앙대학교(2024. 4. 9). 생성형 AI 활용 가이드라인,

<https://www.cau.ac.kr/cms/FR_CON/index.do?MENU_ID=2730> (2024. 12. 12. 최종 접속).

티빙 약관 및 정책 페이지. <<https://www.tving.com/policy/terms>>.

행정안전부 공공지능정책과(2023. 5. 8) “인공지능, 공공부문에서 똑똑하고 안전하게 활용”.
 보도자료
 <https://www.mois.go.kr/frt/bbs/type010/commonSelectBoardArticle.do?bbsId=BBSMS TR_0000000000008&nttId=100278> (2024. 12. 10. 최종 접속).

홍수정(2024. 3. 28). “美법원, ‘AI 가짜 판례’ 인용 변호사에 징직 1년”, 《법률신문》,
 <<https://www.lawtimes.co.kr/news/197090>>.

UNIST(2023, 7). <생성형 AI 활용 가이드라인>. <<https://news.unist.ac.kr/kor/20230817-2/>>.

Harvard Univ.(2023. 7. 13). Initial guidelines for the use of Generative AI tools at Harvard.
 <<https://huit.harvard.edu/news/ai-guidelines>> (2024. 12. 11. 최종 접속).

_____. Generative AI Tool Comparison. <<https://huit.harvard.edu/ai/tools>> (2024. 12. 11. 최종 접속).

_____. Generative Artificial Intelligence. <<https://huit.harvard.edu/ai>> (2024. 12. 11. 최종 접속).

Massachusetts Institute of Technology(2024. 4. 3). Initial guidance for use of Generative AI tools. <<https://ist.mit.edu/ai-guidance>> (2024, 12, 11. 최종 접속).

Stanford Univ.(2023. 2. 16). Generative AI Policy Guidance.
 <<https://communitystandards.stanford.edu/generative-ai-policy-guidance>> (2024. 12. 12. 최종 접속).

Univ. of Oxford. Use of generative AI tools to support learning.
 <<https://www.ox.ac.uk/students/academic/guidance/skills/ai-study>> (2024. 12. 11. 최종 접속).

Yale Univ.(2023. 9. 20). Guidelines for the Use of Generative AI Tools.
 <<https://provost.yale.edu/news/guidelines-use-generative-ai-tools>> (2024. 12. 12. 최종 접속).

● 저 자 소 개 ●

조 성 은

- 릿거스대학교 인터넷커뮤니케이션학 박사
- 현 정보통신정책연구원 연구위원

문 정 욱

- 고려대학교 행정학 박사
- 현 정보통신정책연구원
디지털사회전략연구실 실장

심 홍 진

- 연세대학교 언론학 박사
- 현 정보통신정책연구원 연구위원

김 희 연

- 이화여자대학교 사회학 석사
- 현 정보통신정책연구원 부연구위원

연 제 선

- 성균관대학교 인터랙션사이언스학과 석사
- 현 정보통신정책연구원 연구원

이 대 호

- 현 성균관대학교
소프트웨어융합대학 교수

선 지 원

- 현 한양대학교 법학전문대학원 교수

이 시 직

- 연세대학교 법학 박사 수료
- 현 정보통신정책연구원 부연구위원

김 지 혜

- 서강대학교 국제학 석사
- 현 정보통신정책연구원 전문연구원

박 지 원

- 연세대학교 통계데이터사이언스학 석사
- 현 정보통신정책연구원 연구원

이 지 호

- 경희대학교 문화관광콘텐츠학 석사
- 현 정보통신정책연구원 연구원

이 훈

- 현 경희대학교 미디어학과 교수

정 창 우

- 현 법무법인 광장 변호사

정 낙 원

- 현 서울여자대학교 언론영상학부 교수

정책연구 24-17

지능정보사회 이용자 보호 정책개발

2024년 12월 일 인쇄

2024년 12월 일 발행

발행인 방송통신위원회 위원장

발행처 방송통신위원회

경기도 과천시 관문로 47

Homepage: www.kcc.go.kr

인쇄인 성문화
