발 간 등 록 번 호
11-1620000-000914-01

2022년 특정과제 실태조사
연구용역보고서

인공지능 인권영향평가 도입 방안 연구

2022. 12.

국가인권위원회

인공지능 인권영향평가 도입 방안 연구 (최종보고서)

2022년 국가인권위원회 특정과제 실태조사
연구용역 보고서를 제출합니다.

2022. 12. .

연구수행기관 : 한동대학교산학협력단
연구책임자 : 유승익 (한동대학교 연구교수)
공동연구원 : 김병욱 (해우법률사무소 변호사)
오병일 (진보네트워크센터 대표)
오정미 (법무법인 이공 변호사,
사단법인정보인권연구소 연구위원)
보조연구원 : 안영선 (진보네트워크센터 활동가)

이 보고서는 연구용역수행기관의 결과물로서,
국가인권위원회의 입장과 다를 수 있습니다.

목 차

표 목차

그림 목차

【 요 약 】

1. 인권영향평가 제도 연구

가. 국내외 인권영향평가 현황 및 사례

인권영향평가(Human Rights Impact Assessment, HRIA)는 사업과정, 정책, 입법, 프로젝트 등이 인권에 미치는 영향을 측정하고 평가하는 도구이다. 즉, 인권영향평가는 국가, 지방자치단체, 기업 등 공적·사적 주체들이 시행·추진하는 사업과정이나 정책 등에서 인권에 미치는 부정적 영향을 방지·완화하고, 긍정적인 영향을 미치는 행위를 하도록 장려하기 위해, 법령과 정책 및 사업 등의 계획과 활동이 인권의 실현과 보호에 부합하는지를 평가하고 검토하는 것을 말한다.

1) 해외 인권영향평가의 현황 및 사례

국제연합은 인권영향평가의 원칙과 지침을 형성하는 데 주된 역할을 했다. 특히 국제연합 인권이사회(UN Human Rights Council)의 두 가지 지침이 표준 문서로 많이 거론된다. 2011년의 「기업과 인권에 관한 이행지침(기업과 인권 이행지침)」과 2018년의 「경제개혁 인권영향평가 지침」이 그것이다.

가) 2011년 「기업과 인권에 관한 이행지침」

이 이행지침은 국가의 인권보호의무, 기업의 인권존중책임, 효과적인 구제수단에 대한 접근이라는 세 가지 프레임워크의 실행을 제안했다(보호·존중·구제 프레임워크). 특히 기업경영과 관련하여, 기업은 국제적으로 승인된 모든 인권을 존중해야 원칙을 천명했다. 이를 위해 기업의 인권존중 정책서약, 인권실사 실시, 인권침해에 대한 구제조치를 요구했다. 또한 국가의 인권보호의무와 기업의 인권존중책임을 분리하여 선진국이든 개도국이든 이에 진출한 기업은 해당 국가의 인권보호법제와 독립적으로 인권존중책임을 부담하게 한 데에도 의의를 찾을 수 있다.

인권영향평가와 관련하여 주목되는 부분은 인권실사이다. 이 개념은 이행지침이 천명하는 기업의 인권존중책임의 핵심이라 할 수 있다. 인권영향평가는 인권실사의 핵심도구이다. 인권실사는 ① 인권영향평가의 실시(identify), ② 내부통합(integration), ③ 추적 및 검증(verify), ④ 소통(communication)의 네 가지 과정으로 이루어져 있는데, 그 중심축은 인권영향평가이다. 인권실사는 계속적인 과정으로 기업의 사업활동과 사업맥락에 따라 인권 위험은 달라질 수 있다는 점도 지적되고 있다.

나) 2018년「경제개혁 인권영향평가 지침」

이 지침은 인권에 부합하는 경제개혁 정책의 기본적 조건을 제시하고 있는데, 공공정책에 대한 인권영향평가의 그간의 논의를 종합한 실천 규범이라는 평가를 받고 있다. 이 지침은 22가지 원칙을 5개로 분류하여 제시하고 있는데, ① 경제정책과 인권에 관한 국가 또는 지방정부의 의무, ② 적용가능한 인권 기준, ③ 정책의 구체화, ④ 국가, 국제금융기관, 사적 주체의 기타 의무 그리고 ⑤ 인권영향평가를 내용으로 한다. 이 지침은 국가의 인권영향평가 시행 의무, 인권영향평가의 목적, 평가의 시기, 준수원칙, 평가의 주체 등에 관한 지침을 제공하고 있다.

다) 2010년「인권영향평가 및 관리에 관한 지침」

이 지침은 기업 활동으로 초래되는 인권 위험과 영향을 평가하기 위한 인권영향평가의 원칙과 절차를 단계별로 비교적 상세히 제시하고 있다. 이 지침에서 인권영향평가의 기준은 세계인권선언, 시민적 및 정치적 권리에 관한 국제규약(ICCPR), 경제적, 사회적 및 문화적 권리에 관한 국제규약(ICESCR)을 기본으로 하고 있다. 특히 인권영향평가의 절차를 다음과 같은 7단계로 제시한다. ① 준비, ② 확인, ③ 참여, ④ 사전평가, ⑤ 완화, ⑥ 관리, ⑦ 사후 점검. 또한 기업이 존중해야 하는 '인권'을 35개의 목록으로 제시하고 있다는 점에서 특징적이다.

라) 2020년 덴마크 인권기구의「인권영향평가 가이드 및 도구모음」

이 덴마크 가이드는 기본적으로 유엔 기업과 인권 이행지침에 기반을 두고 있으며, 인권영향평가를 "사업의 맥락에서, 사업 프로젝트 또는 사업활동이 인권에 미치는 부정적

효과를 식별, 이해, 평가, 표명하는 과정”으로 정의한다. 또한 인권영향평가의 10가지 핵심기준을 제시하여, 인권영향평가가 기업 경영뿐만 아니라 공공 정책에도 적용될 수 있도록 원칙과 도구를 세분화·구체화하고 있다. 이 가이드는 차별금지와 같은 인권원칙을 영향평가의 과정에 통합한 인권기반 접근법을 따르고 있으며, ① 계획 및 범위 설정, ② 데이터 수집 및 베이스라인 설정, ③ 영향분석, ④ 영향 완화 및 관리, ⑤ 보고 및 평가의 5단계 절차를 제안하고 있다.

2) 국내 인권영향평가 현황 및 사례

가) 지방자치단체

국내에서 인권영향평가 또는 인권실사는 아직 법률상 규정된 형태로 제도화되지 않았다. 2012년 4월 국가인권위원회 상임위원회는「인권 기본조례 제·개정 권고」와 함께「인권 기본조례 표준안」을 발표하면서 지방자치단체의 조례 제·개정이나 정책 수립에서 인권영향평가 제도를 도입할 근거를 마련하였다. 이후 여러 광역 및 기초지방자치단체에서 인권 관련 조례를 제정하고 인권영향평가를 실시하게 되었다. 지방자치단체는 자치법규(조례, 규칙) 제·개정, 정책·사업 수립·시행 과정에서 시민의 인권침해 및 인권증진 가능성 정도를 사전에 분석·평가하여 행정이 인권증진에 기여할 수 있도록 인권영향평가를 도입하고 있다.

나) 공공기관 및 사기업

공공기관의 경우 인권영향평가(인권실사)는 경영평가제도의 일환으로 간접적 강제가 이루어지고 있다. 여러 공공기관이 인권경영 매뉴얼에 따라 인권정책선언을 수립·공개하고, 인권경영위원회 설립, 인권영향평가 실시 등 인권경영 체계를 구축하여 경영평가를 받고 있다. <공공기관 인권경영 매뉴얼>(2018)은 인권영향평가를 “기관(기업)운영 인권영향평가”와 “주요사업 인권영향평가”로 구분하고 각 시행절차를 제시하고 있다.

법무부의 <기업과 인권 길라잡이>(2021)에서는 인권실사가 대상 및 우선순위 선정, 실재적·잠재적 인권리스크 식별과 평가(인권영향평가), 그 결과를 기업운영과 활동 전반에 반영하고 실천하는 과정, 취해진 조치의 효과성에 대한 모니터링, 그리고 이 모든 절차

에 대한 정보를 공개하는 단계로 구성되며, 모든 단계에서 이해관계자와의 소통과 협력의 중요성을 언급한다.

나. 시사점

국내외의 인권영향평가 제도의 동향에 비추어 우리나라에서 인공지능 인권영향평가의 개발 및 시행에 주는 시사점은 다음과 같다.

첫째, 국내외 동향을 통해 공통적으로 확인할 수 있었던 것은 인권영향평가의 절차 또는 평가단계가 유사하게 수렴되고 있다는 점이다. 인권영향평가의 절차는 대체로 사전준비 단계(계획 및 정보수집), 평가 및 분석 단계, 영향 완화 및 관리 단계, 보고 및 점검 단계 등으로 구성된다. 이 모든 과정에서 이해관계자의 참여가 강조된다.

둘째, 인권실사 및 인권영향평가 제도가 자율적 실시에서 의무적 실시로 제도화·법제화되고 있다는 점이다. 우리나라 인권영향평가는 지방자치단체의 조례에 근거하여 제도화되기 시작하여 경영평가의 일부로 간접강제되는 등 제도적으로 불완전한 형태를 취하고 있다. 인공지능에 대한 인권영향평가에서 명확한 법률적 근거를 마련할 필요성이 있다. 인공지능 기술에 대한 인권영향평가는 평가대상에게 법적 의무를 부과하고 일정한 비용부담을 발생시킨다는 점에서 그 제도화를 위해서는 법률적 근거 마련이 필요하다. 인공지능 인권영향평가의 시행을 위해 기존에 법제화되어 있는 영향평가제도를 활용하는 방안은 일정한 한계를 갖는다. 우리 법제에 난립되어 있는 영향평가제도 가운데 인권영향평가의 성격과 위상이 아직 불분명하며, 규제영향평가 등은 인공지능 인권영향평가 제도를 온전히 실현하기에 여러 한계점을 갖는다. 기존 국가인권위원회법 개정이나 새로운 입법을 통해 인공지능 인권영향평가를 도입하는 방안을 단계적으로 검토할 필요가 있다.

셋째, 인공지능 인권영향평가 도입에서 강조되어야 할 지점은 평가의 방법론이다. 인공지능 인권영향평가는 인공지능이라는 미지의 기술을 대상으로 관련 입법, 정책, 사업, 기술이 인권에 미치는 영향을 측정하고 평가하는 작업이므로 다른 영향평가보다 과학적 방법의 주를 이루는 평가 분야이다. 이는 인공지능 인권영향평가의 성공적 운영에 관건이 되는 인권감수성과 디지털 문해력을 고루 갖춘 전문인력 확보와 밀접히 관련된다.

2. 인공지능 영향평가 사례

인공지능이 사람의 인권과 안전에 미치는 영향에 대한 우려가 커짐에 따라, 그 위험을 예방하고 완화하기 위한 방안으로 영향평가 제도가 주목받고 있다. 최근 유럽연합, 캐나다, 영국, 미국 등 주요 국가들은 인공지능 및 알고리즘에 대한 영향평가 제도를 도입해 왔다. 특히 세계 각국은 공공부문 인공지능과 민간부문 고위험 인공지능이 사람들에게 미치는 부정적 영향에 주목하고 이를 예방적으로 해결하기 위한 방안으로 다양한 인공지능 및 알고리즘 평가를 제안하고 있다.

인공지능 영향평가 제도에 대한 제안은 크게 위험기반 접근(risk-based approach)에서 제안하는 영향평가와 인권기반 접근(rights-based approach)에서 제안하는 영향평가가 각각 논의되어 왔다. 위험기반 접근에서 제안하는 영향평가는 위험 수준이 높아질수록 요구사항을 엄격하게 적용하고 주로 고위험 규제에 초점을 둔다. 한편 인권기반 접근법은 인공지능이 인권에 미치는 위험을 예방하고 완화하기 위하여 인권영향평가 시행을 비롯한 인권실사를 요구해 왔다.

인권영향평가는 그 잠재적 침해의 심각도(severity)를 고려하고 심각성이 높을수록 더욱 신속하고 엄격한 조치를 취하도록 한다. 그런 점에서 인권기반 접근 또한 위험기반 접근을 반영하고 있다고도 볼 수 있다. 그러나 인권영향평가는 특정 인권 항목이 아니라 모든 인권 항목에 대하여 전체적, 포용적, 포괄적 접근방식을 강조하고 영향을 받는 당사자의 참여와 이들에 대한 투명성을 중시하는 국제인권규범을 기준으로 삼는다는 점에서 다른 접근법과 다르며, 기업의 내부적인 위험이 아니라 사람들, 환경, 사회 등 '기업 외부차원'에서 발생하는 부정적 영향으로서 위험을 식별하고 대처하고자 한다는 점에서 여타의 위험평가와 차이가 있다.

가. 인공지능 위험영향평가 사례

1) 유럽연합 위험영향평가

유럽연합은 신기술에 대하여 위험기반접근법을 취해 왔고, 여러 법률에서 신기술을 도

입하는 기관 및 기업에 위험평가의 실시를 요구하여 왔다. 2016년 제정된 <일반개인정보보호규정>은 자동화된 개인정보처리를 비롯한 고위험 개인정보처리에 대하여 '개인정보보호 영향평가'를 적용하여 왔다.

인공지능의 발달로 안전 및 인권에 대한 침해 우려가 커짐에 따라 유럽연합 집행위원회는 2019년 <신뢰할 수 있는 인공지능 평가 목록>을 제안하고 2021년 제안한 <인공지능법(안)>에서 고위험 인공지능에 대한 위험평가를 제도화하였다. 고위험 인공지능 위험관리 시스템의 경우, 고위험 인공지능 시스템의 수명주기 전반에 걸쳐 지속적으로 운영되고 정기적·체계적으로 갱신을 반복해야 하며, (a) 고위험 인공지능 시스템과 관련된 알려지고 예측가능한 위험의 식별 및 분석, (b) 고위험 인공지능 시스템을 합리적으로 예측 가능한 오남용 조건 하에서 원래 목적으로 사용할 때 발생할 수 있는 위험의 추정 및 평가, (c) 출시 후 모니터링 시스템에서 수집한 데이터의 분석에 근거한 기타 위험 발생가능성의 평가, (d) 적합한 위험 관리 수단의 채택으로 구성된다. 특히 설계와 개발을 통해 최대한 위험을 제거 또는 완화하고, 제거할 수 없는 위험에 대해서는 완화 및 통제 조치를 시행하는 등 적합한 위험 관리 수단을 채택한 후에도 잔여 위험이 남아 있을 경우, 이를 허용 가능한 수준으로 보장하고 사용자에게 통지해야 한다. 고위험 인공지능 시스템은 테스트를 통해 원래 목적에 일치하고 요구사항을 준수하는지 여부를 확인해야 하며, 이러한 테스트는 개발 과정에서 임의의 시점에 수행하되, 어떠한 경우에도 출시 또는 서비스 개시 전에 수행되어야 한다.

한편, 2022년 11월 16일 발효한 유럽연합 <디지털서비스법>은 체계적 위험이 우려되는 대규모온라인플랫폼의 알고리즘 등에 대하여 위험평가를 실시하도록 하였다.

2) 캐나다 알고리즘영향평가

캐나다 정부는 2019년 연방정부의 조달 정책 및 제도를 소관하는 재정위원회의 훈령으로 「자동화된 의사결정 훈령」을 제정하여 시행 중이다. 이 훈령은 인공지능 시스템에 대한 요구사항을 법규화하고 있는데, 평가된 위험 수준이 높을수록 적용되는 요구사항도 높아진다. 공공기관에서 도입하는 자동화된 의사결정 시스템은 요구사항의 기반이 되는 위험 수준을 측정하기 위하여 알고리즘영향평가를 의무적으로 실시하여야 한다. 실

시 시기는 프로젝트 설계 단계 초기에 우선 실시하고, 시스템의 생산 전에도 두 번째로 실시하여 요구사항이 구축된 시스템에 반영되었는지 확인하도록 하였다. 두 번째 평가 결과는 일반 접근이 가능한 형식으로 온라인에 공개하여야 한다. 시스템의 기능 또는 범위가 변경되면 평가를 갱신하여야 한다.

평가도구인 캐나다 알고리즘영향평가는 자동화된 의사결정 시스템의 위험 및 완화 정도에 대하여 묻고 영향 수준을 판단하는 질문지로 구성되어 있다. 위험 점수는 48개의 위험과 33개의 완화 조치에 대한 질의와 그 답변에 기반하여 산출되며, 완화 점수가 80% 이상 도달하면 위험 점수를 15% 차감한다. 위험을 측정하는 질의는 프로젝트, 시스템, 알고리즘, 의사결정, 영향, 데이터에 대한 다방면 질의로 이루어지며, 완화에 대한 질의는 이해관계자 협의, 데이터 품질, 절차적 공정성, 개인정보보호에 대한 내용들이다.

훈령은 자동화된 의사결정 시스템의 책임 있는 사용을 지원하기 위해 범정부 차원에서 또는 타부문 관할 기관과 교류하고 참여하도록 하였다. 예를 들어 평가 중인 알고리즘 시스템이 개인정보와 관련된 경우, 개인정보보호 영향평가를 수행하거나 수행한 적이 있는지 묻는 등 개인정보 보호법의 원칙과 규정을 참조하였다.

이 훈령은 외부 대상으로 추천이나 행정적 의사결정을 하기 위하여 알고리즘 시스템, 도구, 통계적 모델을 개발하거나 조달하는 연방정부 기관을 그 적용 범위로 한다. 평가는 자동화된 의사결정 시스템을 생산하려는 해당 기관이 온라인에서 직접 수행한다. 이 훈령은 캐나다 정부 재정위원회가 소관하지만, 평가 기준이나 점수 산정 등 알고리즘영향평가의 개발 및 유지관리는 캐나다 최고 정보 책임자(CIO)가 담당한다. 훈령을 준수하지 않는 경우 재정위원회의 조치 대상이 될 수 있다.

3) 영국 인공지능 영향평가

가) 인공지능 조달지침

영국 정부는 2020년 6월 인공지능 조달지침을 발표하고 두 가지 방식의 평가를 요구하였다. 우선 조달 절차 개시 전에는 데이터 평가를 실시하고, 조달 절차 개시 단계에서 인공지능 배치의 편익과 위험에 대하여 영향평가를 실시하여야 한다.

우선 조달 기관은 데이터 평가를 통하여 △조달 절차의 개시 단계부터 데이터 거버넌

스 메커니즘이 가동될 수 있도록 확보하고, △프로젝트에 관련 데이터를 사용할 수 있는지 여부를 평가하며, △시장에 출시하기 전에 데이터 내부의 결함 및 편향 가능성을 해결할 수 있어야 한다. 데이터 문제를 직접 해결할 수 없는 경우 이를 해결하기 위한 계획을 수립해야 한다. 더불어 △조달 계획 및 후속 프로젝트를 위해 공급업체와 데이터를 공유할 것인지 여부 및 방법을 정의할 것 또한 요구한다. 데이터에 대한 철저한 평가가 어려운 것으로 드러나거나 이루어지지 않은 경우, 인공지능 시스템이 의사결정의 기반으로 사용할 데이터에 대해 종합적인 점검을 실시할 것을 입찰공고 요구사항에 포함하여야 한다. 입찰 공고는 데이터를 덜 침해적으로 사용하거나 덜 민감한 데이터셋을 이용하여 동일하거나 유사한 결과를 달성하는 혁신적인 기술 접근법을 장려해야 한다.

다음으로 조달 절차 개시 단계에서 인공지능 영향평가를 수행하고, 중간 평가 결과가 조달에 반영되는지 확인하여야 한다. 주요 의사결정 단계에서는 평가 결과를 재차 살펴보아야 한다. 즉, 인공지능 영향평가는 프로젝트 설계 단계에서 실시하고 이후 솔루션 설계 및 조달 절차에서 확인된 위험성의 완화를 추구해야 하며, 장래 구현될 인공지능 시스템에 대한 완전한 평가가 사실상 불가능하기 때문에 반복적으로 점검하여야 한다. 의사결정 또는 인공지능 시스템 설계에 상당한 변화가 있을 때마다 영향평가를 검토하여야 한다.

조달지침이 제시하는 인공지능 영향평가의 항목은 6가지로서, ①인공지능 시스템에 대한 사용자 요구사항과 그 공익, ②인공지능 시스템의 인적·사회경제적 영향, 즉 인공지능이 사회적 가치 편익을 제공할 수 있도록 보장하는지, ③기존의 기술적, 절차적 환경에 미치는 결과, ④데이터 품질 및 부정확성 또는 편향성, ⑤의도하지 않은 결과가 나올 가능성, ⑥지속적인 지원 및 유지보수 요구사항 등 전체 생애주기에 대한 비용적 고려 등이다.

나) NMIP 알고리즘영향평가

에이다 러브레이스 연구소와 국민 보건 서비스 NHS는 어떤 인공지능 시스템이 NHS AI Lab의 국가의료이미지플랫폼(NMIP)의 데이터를 활용하고자 할 때 그 영향을 평가하는 도구를 개발하였다.

NMIP 알고리즘영향평가는 7개의 절차로 이루어진다. ①성찰적 수행 단계는 NMIP의 데

이터 액세스 위원회(DAC)에 데이터 접근 신청서를 제출하려는 신청자팀이 스스로 템플릿을 이용하여 프로젝트에 대한 설명을 작성하고 시스템이 미치는 영향을 평가하는 일련의 질의와 답변을 거치도록 하였다. 의사결정기구인 DAC는 사회과학, 생의학, 컴퓨터 과학, 법학 등 학계 전문가, 환자단체 대표 등 11명 이상으로 구성된다. ②신청서 필터링을 통해 심사기준을 충족한 신청자는 다양한 이해관계자 패널들과 ③참여 워크숍에 참여하여 프로젝트의 잠재적인 영향에 대해 논의하고 NHS는 이에 대한 보고서를 작성한다. 이해관계자 패널은 연령, 성별, 지역, 민족적 배경, 사회경제적 배경, 건강 상태 또는 치료 접근성에 걸쳐 알고리즘의 영향을 받을 수 있는 인구의 다양성을 반영하는 8-12명으로 구성된다. 참여 워크숍 후 신청자는 워크숍 결과를 ④종합하여 ①에서 작성하였던 템플릿을 업데이트한다. ⑤데이터 접근여부 결정을 위하여 업데이트된 템플릿과 참여 워크숍 보고서를 DAC에 제출하면, DAC는 NMIP 알고리즘영향평가의 내용과 기타 자료를 검토하여, NMIP 데이터 접근 여부를 결정한다. NMIP 알고리즘영향평가 결과물은 ⑥공개되며, 2년의 주기 또는 DAC 재량에 따라 ⑦반복적으로 실시될 수 있다. 인공지능 시스템이 제품의 기능, 범위 및 애플리케이션의 변경, 사용자 기반의 변화 등 중대한 변경이 이루어질 경우 영향평가를 다시 수행해야 한다.

템플릿은 4개의 섹션으로 구성되며, 첫째, 프로젝트에 대한 정보 섹션에서는 시스템 및 모델의 입력 및 결과물에 대한 세부정보(데이터 소스 등) 및 시스템의 영향을 받는 이해관계자에 대한 정보를 작성한다. 둘째, 윤리적 고려사항에 대한 섹션은 특정 커뮤니티에 대한 차별 가능성, 동의와 자율성, 감시 위험, 영향을 받는 당사자가 결과에 이의를 제기할 수 있는 방법, 의도하지 않은 오류 혹은 의도적인 오용의 가능성 등을 평가한다. 셋째, 영향 식별 및 시나리오 섹션은 최상의 시나리오와 성공적 운영을 위한 요구조건, 과제나 장애물은 무엇인지, 최악의 시나리오의 경우 시스템이 의도된 대로 작동하지 않을 때 발생할 수 있는 상황은 무엇인지 등을 분석하도록 한다. 넷째, 잠재적 피해 분석 섹션은 이해관계자의 잠재적 피해 가능성과 완화조치를 분석한다. 피해를 분석할 때에는 피해의 중요도, 긴급성, 완화조치의 어려움, 탐지가능성을 고려한다.

NMIP 알고리즘영향평가는 기존의 규제 메커니즘과 대체되거나 중복되는 것이 아니라 상호 보완적이라고 설명한다. 즉, 규제 대상이 되는 기관들은 NMIP 알고리즘영향평가와 무관하게 개인정보보호 영향평가, 의약품 기기 위험 분류 등의 규제를 준수해야 한다.

영향평가 모델이 성공적이려면 독립적인 평가자가 중요성하다. 설명 책임(책무성)이 중요할 경우 평가의 독립성이, 성찰성에 중점을 둘 경우에는 시스템 개발자의 자체 평가를 우선시할 수 있다. NMIP의 사례에서는 개발자를 위한 성찰적인 절차를 허용하고, DAC가 독립 평가자로서 이를 검토하는 절차를 둠으로써 두 가지 관심사를 모두 포착하려 하고 있다.

4) 미국 알고리즘영향평가

2022년 10월 미국 백악관 과학기술정책국은 <인공지능 권리장전 청사진>을 발표하면서 인공지능이 준수하여야 할 5가지 원칙으로 안전하고 효과적인 시스템, 알고리즘 차별로부터 보호, 개인정보 보호, 통지 및 설명, 인간 대안·검토 및 대체를 제시하면서 각 분야 위험을 식별하고 완화하기 위하여 영향평가의 수행과 공개를 강조하였다.

2022년 2월 3일 미국 하원과 상원에 함께 발의된 <알고리즘 책무성법(안)>은 연방거래위원회(FTC)가 소관하는 일정 규모 이상의 기업들을 대상으로 자동화된 의사결정 시스템 또는 증강된 중요 의사결정 프로세스 및 이들이 소비자에게 미치는 영향에 대하여 지속적으로 연구·점검하는 '영향평가'를 실시하도록 의무화하고 FTC가 감독하도록 하였다. 여기서 '중요 의사결정'이란 교육평가, 고용·근로자 관리·자영업, 전기·난방·수도·통신·교통 등 필수 설비, 가족서비스, 금융서비스, 보건의료서비스, 주택서비스, 법률서비스 등이다.

대상 기업은 수행된 영향평가 관련 문서를 시스템 또는 프로세스 배치 후 3년 이상 보관하여야 하며, 영향평가에 대한 요약보고서를 매년 FTC에 제출하는 한편, 신규 시스템 또는 프로세스의 경우 초기 요약보고서를 그 배치 전에 제출하여야 한다. 대상 기업은 영향평가 수행시 관련 내부 이해관계자(직원, 윤리 팀 및 담당 기술팀 등) 및 독립적인 외부 이해관계자(영향을 받는 집단의 옹호자나 대표, 시민 사회 및 인권단체, 기술 전문가 등)와 필요에 따라 수시로 의미 있는 협의(참여 설계, 독립 감사 또는 피드백 요청 및 통합)를 하여야 한다. 소비자의 삶에 법적 또는 유사하게 중대한 영향을 미치는 물질적·부정적 영향이 나타나는 경우 프로세스에 의해 발생하는 모든 영향을 시기적절한 방식으로 제거하거나 완화하기 위해 노력하여야 한다.

영향평가 요구사항은 ①기존 프로세스에 대한 검토 ②이해관계자 협의 ③개인정보보

호 테스트 및 검토 ④현재 및 과거 성능에 대한 지속적인 테스트 및 검토 ⑤직원에 대한 지속적인 교육훈련 ⑥시스템 또는 프로세스의 특정한 사용 및 적용에 대한 보호막이나 한정의 필요성과 개발 가능성 ⑦개발, 테스트, 유지 관리, 갱신하는 데 사용되는 데이터 및 기타 입력 정보에 대한 최신 문서의 유지 관리 및 보관 ⑧소비자의 권리 ⑨소비자에게 미치는 중대한 부정적 영향 가능성의 식별 및 적용가능한 완화 전략 ⑩개발 및 배치 절차에 대하여 진행 중인 문서화 ⑪개선이 필요한 기능, 도구, 표준, 데이터셋, 보안 프로토콜, 이해관계자 참여 및 기타 자원 ⑫미준수 영향평가 요구사항 및 미준수 근거 등이다.

나. 인공지능 인권영향평가 사례

1) 유엔 인권규범과 인공지능 인권실사

유엔 인권기구들은 인공지능 등 신기술이 인권에 미치는 부정적인 영향을 식별·방지·완화하기 위하여 인권실사의 시행을 권고하여 왔으며, 인권영향평가는 인권실사의 유용한 도구로서 인권에 미치는 부정적 영향을 식별하고 대처하는 데 도움이 된다고 여러 차례 강조하였다. 특히 유엔 인권최고대표는 2021년 〈디지털 시대 프라이버시권〉 보고서에서 국가와 기업에 대하여 “인공지능 시스템의 설계, 개발, 배치, 판매, 구입, 운영의 수명 주기 전반에 걸쳐 체계적으로 인권실사를 수행”할 것을 권고하고, “그 인권실사의 핵심 요소는 정례적이고 포괄적인 인권영향평가여야 한다”고 강조하였다. 유엔인권이사회는 2021년 10월 13일 〈디지털시대 프라이버시권 결의〉에서 “국가 및 적용대상 기업이 설계, 개발, 배치, 판매 또는 구입 및 운영하는 인공지능 시스템의 수명주기 전반에 걸쳐 인권실사를 실시할 것을 권장”하는 등 인공지능에 대한 인권실사를 강하게 요구하였다.

2) 유럽평의회 권고와 인권·민주주의·법치 영향평가

유럽평의회 인권위원장은 2019년 보고서에서 인공지능이 인권에 미치는 부정적인 영향을 예방하고 완화하기 위한 첫번째 방안으로 인권영향평가를 권고하였다. 이후 유럽평

의회는 2020년 4월 8일 알고리즘 시스템의 인권영향에 대한 각료위원회 권고 CM/Rec(2020)1 와 부록 <알고리즘 시스템의 인권영향에 대한 대응 지침>을 채택하고, 특히 인권영향평가를 국가와 민간이 의무적으로 취하여야 할 예방적 조치로 보고 상세한 요구사항을 설명하였다.

유럽평의회 인공지능 특별위원회(CAHAI)는 2021년 5월 21일 <인공지능 시스템에 대한 인권・민주주의・법치 영향평가(HRDRIA)> 초안 문서를 발표하였다. 공공 및 민간 모두에 적용되는 HRDRIA는 '초기 평가'에서 인권 위험이 높게 나왔을 경우 수행할 것을 제안하고 있다. 모든 인공지능 시스템을 대상으로 수행하는 것은 시간과 비용이 과도하게 들기 때문이다. 초기 평가는 애플리케이션의 규모, 형태와 목적, 체계적으로 인간과 상호작용하는 정도, 특별히 위험한 사용 사례(예를 들어 얼굴인식, 딥페이크, 소셜네트워크) 등이 고려될 수 있을 것이다.

HRDRIA는 4개의 절차로 이루어진다. ①인공지능 시스템에 의해 부정적 영향을 받을 가능성이 있는 관련 권리를 식별하고, ②해당 권리에 대한 영향평가를 시행한다. 영향평가는 기술적, 비기술적 측면을 모두 포함하여, 기술적 분석은 인공지능의 설명가능성, 투명성, 사이버보안, 보호조치 등에 초점을 맞춘다. 비기술적 분석은 시스템이 작동하는 사회-기술적 환경, 인공지능 보급 및 사용에 필요한 역량과 기술을 분석하고, 인공지능 보급의 위험성에 대한 식별, 해결, 추적 또한 검토한다. 더불어 민주주의 및 법치의 대체변수에 해당하는 기본권에 대한 영향도 분석한다. ③이해관계자 참여나 불만처리 메커니즘 등 잠재적 위험을 완화하는데 도움이 될 거버넌스 메커니즘이 있는지 검토한다. ④완화조치들을 지속적으로 점검하고 지속적인 영향평가 실시를 검토한다. HRDRIA는 개발자뿐만 아니라 판매자, 조달자, 배포자에 의해서도 수행되어야 한다. 인권에 미치는 영향의 심각성, 규모, 회복불가능성 등에 따라 적절한 규모의 이해관계자가 평가 절차에 참여하여야 한다.

3) 덴마크 디지털활동 인권영향평가

덴마크 국가인권기구는 2020년 <디지털활동 인권영향평가 지침>에서 우선 인권실사가 기업 활동 전반에 걸쳐 시행되어야 하는 반복적인 과정이라는 점에서 인권영향평가 또한 언제, 어떻게 실시할지는 사업별로 특유하다고 설명한다. 얼마나 정밀한 평가를 수행해

야 하는지 역시 규격화할 수 없으며, 식별된 위험, 그 심각도, 잠재적 또는 실제 영향에 대한 회사의 자원 및 참여, 기타 다양한 요인에 따라 달라진다. 인권영향평가를 실시한 후로도 제품 및 서비스의 규모, 범위, 사용 또는 적용이 변경될 때에는 재평가되어야 한다.

덴마크 인권영향평가는 5개의 절차로 이루어진다. ①계획 및 범위 설정 단계는 디지털 활동의 유형, 인권적 맥락, 관련 이해관계자, 평가 결과의 활용처와 관련한 정보를 검토하여 인권영향평가의 과업을 정의한다. 기관과 담당자는 이 단계에서 피평가기관으로부터 독립적인 인권영향평가팀의 구성과 이해관계자의 참여에 대한 결정을 내린다. ②데이터 수집 및 상황 맥락 분석 단계에서는 사용자, 영향을 받는 권리주체, 특히 취약 집단의 인권 향유에 대한 기본 데이터를 이해관계자로부터 수집한다. 이렇게 수집된 데이터를 통해 평가팀은 인권 향유의 현재 상태에 대한 맥락 분석을 수행할 수 있다. 맥락 분석은 인권에 미치는 실제 영향을 식별하고 향후 영향을 더 잘 예측하는 데 도움이 된다. ③영향 분석 단계는 실제적 또는 잠재적인 인권영향을 식별하고 심각도를 평가하기 위해 앞서 수집된 데이터를 분석한다. 분석에는 국제 인권 기준, 사업 비교, 이해관계자 참여 결과 등을 활용한다. 인권영향 분석은 '즉각적'으로 보이는 영향뿐 아니라 사업이 야기하고 기여했거나 그럴 수 있는 모든 영향, 직접적으로 관련된 영향을 고려하여야 한다. 영향 분석에는 영향의 범위, 규모 및 회복 불가능성을 고려하여 영향의 '심각도'를 평가하는 과정이 포함되어야 한다. 인권영향의 심각도가 높을수록 신속하고 엄격한 조치가 필요하다. 이때 영향을 경험하였거나 경험할 수 있는 사람의 관점을 고려해야 한다. 무엇보다 인권영향평가가 인권 존중에 기여하기 위해서는 부정적인 인권영향을 우선적으로 식별하고 해결하는 데 중점을 두어야 한다. ④영향 예방, 완화 및 구제 단계는 기관, 평가팀 및 이해관계자는 협력하여 부정적인 인권영향을 예방, 완화 및 개선하기 위한 계획을 수립한다. 모든 인권영향을 해결하는 것이 목표이며 가장 심각한 영향을 우선적으로 고려해야 한다. 식별된 영향에 대한 조치는 주로 부정적인 인권영향을 방지하고 감소시키는 데 중점을 두어야 한다. 영향 관리 계획이 수립되면, 후속 조치를 시행하여 식별된 영향을 효과적으로 해결하는 것이 중요하다. 이때 지속적인 모니터링은 영향 완화 조치가 효과적인지 여부에 대한 정보를 제공하고 그렇지 않은 경우 필요한 조정을 수행할 수 있도록 하며, 예상치 못한 영향도 식별할 수 있도록 한다. 구제책에 대한 접근 또한

영향 관리의 핵심 요소인 만큼, 고충 처리 체계를 마련하여야 한다. 고충 처리 체계를 통하여 인권영향평가 및 그 모니터링과 관련된 고충을 처리하고 디지털활동의 인권영향을 계속 식별할 필요가 있다. ⑤보고, 모니터링 및 점검 단계에서는 영향평가 보고서를 작성하고 공개하여 권리주체, 의무주체 및 기타 관련 당사자가 이용할 수 있도록 한다. 인권영향평가는 그 효과성 평가 및 지속적인 개선 조치까지 포함하여야 한다. 정례적인 검토는 인권영향평가 후 발생할 수 있는 문제를 해결하는 데 도움이 된다. 마지막으로 이해관계자 참여는 인권영향평가의 모든 단계에서 공통적으로 요구되는 구성요소이며, 이때 이해관계자는 권리주체, 의무주체 및 기타 이해관계자를 아우르는 개념이다.

덴마크 인권영향평가의 적용 대상은 공공기관과 민간의 개발 사업자 또는 구매 사업자이며 인권실사의 일부로서 인권영향평가를 자율적으로 실시하도록 하였다. 다만 인권영향평가의 실시를 담당하는 팀이 기관과 독립적일 때 예방 및 완화 조치의 관측과 권장사항 도출의 정당성을 보장하고 기관의 인권 담당 직원을 적절하게 지원할 수 있다. 인권영향평가팀이 전적으로 내부 직원으로 구성될 경우 평가의 독립성을 제한하게 되지만, 외부인으로는 기관 내부에 대한 지식과 전문성이 결여될 수밖에 없다. 따라서 평가팀, 회사 임직원, 기타 이해관계자가 함께 평가를 위한 팀을 구성하는 것이 바람직하다.

4) 네덜란드 기본권 알고리즘영향평가

네덜란드 내무부에서 마련한 기본권 알고리즘영향평가는 알고리즘 시스템의 개발 또는 조달을 검토 중인 공공기관이 계획의 초기 단계에서 인권과 관련된 쟁점을 검토할 수 있는 질의를 제공한다. 이 질의는 공공기관 의사결정자가 4개의 단계를 거쳐 대화 지향적이고 질적인 접근 방식으로 내부와 외부의 이해관계자와 협의하고 알고리즘 시스템의 개발 및 조달의 진행 여부와 방법에 대하여 적절한 결정을 내릴 수 있도록 지원한다. 기본권 알고리즘영향평가의 답변과 개인정보보호 영향평가 등 타 평가의 답변은 상호 활용할 수 있다.

1부는 준비 단계로 알고리즘이 사용되는 이유와 그 효과가 무엇인지 판단한다. 2부는 알고리즘을 개발하는 데 어떤 데이터가 사용되는지, 알고리즘이 어떤 형태여야 하는지를 결정한다. 특히 데이터에 대해서는 특정 유형의 데이터 및 데이터 소스의 사용에 대하여 질의하고, 알고리즘에 대해서는 알고리즘의 작동 및 투명성에 관해 질의한다. 3부는 출

력, 구현 및 감독 단계로, 알고리즘을 사용하는 방법에 관한 것이다. 즉, 알고리즘이 생성하는 결과물이 무엇인지, 그 결과물이 정책 또는 의사결정에서 어떤 역할을 할 수 있는지, 이를 어떻게 감독할 수 있는지 질의한다. 마지막으로 4부는 사용할 알고리즘이 기본권에 영향을 미치는지 여부를 파악하고, 기본권 행사에 대한 간섭을 방지하거나 완화할 수 있는지, 기본권 간섭이 수용 가능한지 등을 검토한다. 특히 기본권 알고리즘영향평가는 심각도를 세 가지 등급으로 구분하였으며, 심각도 등급에 따라 부과되는 요구사항 또한 차등 적용될 수 있다고 보았다. 기본권 침해가 기본권의 핵심에서 멀어질수록 침해가 덜 심각하고, 보다 온건한 실사 및 정당화 요구가 이루어진다.

5) 주요 빅테크 기업의 인권영향평가

구글, 페이스북, 마이크로소프트 등 주요 빅테크 기업들은 자사의 인공지능 혹은 알고리즘에 대해 인권영향평가를 수행한 바 있다.

가) 구글의 유명인 인식 API에 대한 인권영향평가

구글은 기업 고객이 구글의 유명인(celebrity) 이미지 데이터베이스를 사용하여 자신의 콘텐츠에서 프레임이나 씬 단위로 유명인을 식별할 수 있는 API를 개발하면서 인권영향평가를 수행하였다.

영향평가를 의뢰받은 BSR은 유엔 기업과 인권 이행지침에 기반한 방법으로 인권영향평가를 수행했다고 한다. 즉, 인권영향평가 과정에서 영향을 받는 이해관계자와 협의를 하였고, 독립적인 전문가와 대화하였으며, 취약성이 큰 그룹에 특별한 관심을 기울였다. 평가 과정에서 구글 클라우드 AI의 API 제품팀 및 AI 원칙팀과 협력하였다. 구글이 재정지원을 하였지만, BSR은 보고서 내용에 대해 통제권을 보유하여 독립성을 보장받았다고 한다.

나) 페이스북의 국가별 인권영향평가

2018년 11월 5일, 페이스북은 미얀마에서 페이스북의 역할에 대한 인권영향평가 보고서를 발표하였다. 이 인권영향평가 역시 BSR에 의해 수행되었다. 이 보고서는 2018년 이

전 페이스북이 자신의 플랫폼이 분열과 폭력의 조장을 방지하는데 충분한 역할을 하지 못했다고 결론을 내렸다. 2018년에는 페이스북이 미얀마에서의 페이스북의 남용을 막기 위해 인력, 기술, 파트너쉽에 더 많은 투자를 했으며 보고서는 이러한 시정 조치들을 인정한다고 밝혔다. 보고서에서 BSR은 5가지 권고안을 제시했는데, 첫째, 거버넌스 및 책임성 구조에 기반할 것, 둘째, 콘텐츠 정책의 집행을 증진할 것, 셋째, 지역 이해관계자의 참여를 강화할 것, 넷째, 규제 개혁을 옹호할 것, 다섯째, 미래를 준비할 것 등이다.

그러나 하버드 케네디스쿨의 인권정책을 위한 카센터의 한 연구자는 페이스북의 미얀마에서의 인권영향평가가 미얀마에서의 인권 침해를 야기하거나 기여한 것에 대한 페이스북의 책임을 제대로 다루지 않았다고 비판했다. 이 인권영향평가가 페이스북 뉴스피드 알고리즘의 인권영향에 대해 다루지 않았고, 미얀마 맥락의 핵심적 요소들이 미얀마 뉴스 피드 운영 결정과 관련되는지에 대해 다루지 않았다는 것이다. 2022년 7월, 메타는 2020-2021년 활동을 다룬 메타 인권보고서를 발행했는데, 여기서는 인도 인권영향평가 보고서를 은폐했다는 비판을 받았다. 전체 보고서 뿐만 아니라 보고서의 권고가 무엇인지도 자세한 내용이 공개되지 않았기 때문이다.

메타의 인권영향평가 사례는, 인권영향평가가 자칫하면 기업으로 하여금 인권 보호 책임을 충족하는 외양만 갖추게 할 뿐, 실질적으로는 그 책임을 회피하는 수단으로 전락할 수 있음을 보여준다. 또한, 인권영향평가의 독립적 수행과 보고서 공개를 통한 투명성 확보가 인권영향평가의 신뢰성을 위해 매우 중요하다는 점을 알 수 있다.

다) 마이크로소프트의 책임있는 인공지능 인권영향평가 가이드

2022년 6월, 마이크로소프트(MS)는 인권영향평가 템플릿이 포함된 <책임있는 인공지능 영향평가 가이드>를 발간하였다. 이 가이드는 서로 다른 전문성을 가진 사람으로 평가팀을 구성하고 템플릿에 구성원의 토론 내용을 기록하도록 하고 있다. 템플릿의 내용은 이후 잠재적인 평가자가 검토하는 관점에서 작성된다.

이 가이드는 ①프로젝트 개요, ②의도된 사용, ③부정적 영향, ④데이터 요구사항, ⑤영향 요약의 5개 섹션으로 구성되어 있다.

다. 국내 인공지능 기준과 인권영향평가

1) 인공지능 자율점검 기준

인공지능이 사회에 미치는 영향에 대한 우려가 커짐에 따라 우리나라에서도 개인정보보호위원회, 과학기술정보통신부, 금융위원회, 서울특별시교육청 등 중앙부처 및 지방자치단체가 인공지능의 분야별 위험을 점검하기 위한 기준 및 도구를 보급하여 왔다. 현재까지 이들 도구들은 자율점검 절차로 사용되고 있다.

국내에서 인공지능에 대한 구속력 있는 영향평가 제도가 도입된다면 위의 기준들이 그 평가도구에 반영될 수 있을 것이다. 다만, 위 도구들은 인공지능이 미치는 영향 중에 각 부처 소관사항을 살펴볼 뿐, 인권에 미치는 부정적인 영향을 식별하고 방지 및 완화하는 것을 주요 목적이나 내용으로 삼고 있지 않다.

2) 인권경영과 인공지능 인권영향평가

우리나라에서 인권실사 및 인권영향평가는 아직까지 법률로 규정되어 있지 않다. 다만 정부가 2021년 12월 30일 국회에 발의한 인권정책기본법안에서 '제5장 기업과 인권'은 유엔 기업과 인권 이행지침의 구조와 개념을 수용하였다.

현재 우리나라 공공기관과 기업에 대한 인권영향평가는 인권실사를 의무화해 온 국제기준의 압력을 받은 경영평가제도를 통하여 간접적으로 의무화되고 있다고 볼 수 있다. 특히 상장기업 우선으로 환경・사회・지배구조(ESG) 공시가 추진되면서 기업과 그 사업에 대한 인권영향평가가 도입되고 있다. 기획재정부는 2018년도 <공공기관 경영평가편람>에서 윤리경영 지표에 인권 항목을 포함한 데 이어, 2022년 2월 4일 인권경영 항목을 독립적으로 신설 및 공시하도록 하였다. 이러한 배경에서 공공기관과 공기업에도 인권영향평가의 실시 및 공개가 확산되어 왔다.

국가인권위원회와 각 정부부처는 인권경영에서 인권영향을 평가하기 위한 도구들을 개발하고 보급하여 왔다. 법무부는 2019년 5월 <인권경영 표준지침(안)>을 발표하였고, 산업통상자원부는 2021년 12월 <K-ESG 가이드라인 v1.0>을 발표하였다. 특히 국가인권위원회는 2014년 <인권경영 가이드라인 및 체크리스트>와 2018년 <공공기관 인권경영

매뉴얼>을 발표하고 공공기관장들에게 그 적용 및 활용 등을 권고하여 왔고, 현재 대부분의 공공기관은 이 기준과 도구를 참고하여 인권경영체계를 구축하고 있다.

국가인권위원회 <공공기관 인권경영 매뉴얼>이 권고하고 있는 인권영향평가 추진체계 및 도구에 대하여 살펴보면 다음과 같다. 우선 인권경영 대상 기관은 인권경영 담당 부서와 담당자를 지정하고 인권경영위원회를 구성하여야 한다. 인권경영 담당부서는 인권영향평가 실시 계획 및 체크리스트를 마련하는 한편 기관 내부 교육을 실시하는 등 인권영향평가 절차를 주무하는 역할을 한다. 인권영향평가를 실시하는 인권경영위원회는 임직원, 노동조합, 공급망, 지역주민, 고객, 인권전문가 등 기관 내외부 이해관계자로 구성된다.

인권경영 평가도구로는 국가인권위원회 가이드라인 및 체크리스트 지표 등을 기반으로 각 기관의 실정에 맞는 체크리스트를 마련할 것이 권장된다. 사업 인권영향평가의 경우 인권경영 담당부서가 먼저 사업부서와 이해관계자로부터 다양한 정보와 의견을 수렴하여 대상 사업의 실제적·잠재적 인권위험을 분석하고, 이렇게 분석한 위험을 바탕으로 지표를 선정하고 인권영향평가 체크리스트를 구성한다.

특히 국가인권위원회는 2022년 <인권경영 보고 및 평가 지침>에서 인권경영 보고가 '주요 인권이슈'를 반드시 명시할 것을 요구하며, '중대성 평가'를 통해 기관 및 기업이 해당년도에 대응하기로 결정한 주요 인권이슈가 인권영향평가의 최종결과물로서 도출되어야 한다고 강조하였다.

한편, 국가인권위원회는 2022년 5월 17일 <인공지능 개발과 활용에 관한 인권 가이드라인>을 공개하였다. 국내에 도입되는 인공지능 인권영향평가는 국가인권위원회 가이드라인에서 제시하는 원칙에 부합하는 방향으로 도입되는 것이 바람직할 것이다. 인권영향평가를 실시하는 기관과 기업들은 국가인권위원회 등에서 보급된 평가도구들을 그대로 사용하거나 최적화하여 사용해 왔다. 각 기관이나 기업이 인공지능 사업에 대한 인권영향평가를 실시할 때 본 연구에서 개발한 도구를 비롯하여 국내외에서 제안되어 온 다양한 인공지능 평가 기준 및 도구들을 여타의 인권영향평가도구들과 결합하고 최적화하여 사용할 수 있을 것이다.

다만 최근 세계 각국은 자발적인 보고 의무를 중심으로 한 현행 인권실사제도가 효과적이지 못하고 관행을 변화시키지 못한다는 비판적 문제의식 속에 인권실사 실시 의무를

법적으로 부과하기 위해 노력하고 있다. 우리나라에서도 인권실사의 실질화 또는 의무화에 대한 논의가 활발하게 이루어지고 있는 만큼, 인공지능 인권영향평가 역시 인권실사의 제도화 흐름과 조응할 필요가 있다.

3. 인공지능 영향평가 도입 방안

가. 제도적 형식의 측면

인공지능 인권영향평가는 고유의 목적을 실현할 수 있도록 기존의 다른 사회영향평가에 결합시키기보다 단독 형태로 명확한 법률적 근거를 마련할 필요가 있으나, 현 단계에서도 제도 시행 자체는 얼마든지 가능하다. 인공지능 기술을 활용한 사업 또는 정책이 인권에 미치는 영향을 적절히 식별하고, 관리하기 위하여 공공기관이 직접 개발하거나 조달을 통해 활용하는 모든 인공지능과 민간에 도입되는 위험성이 높은 인공지능에 대하여 사전에 인권영향평가를 수행하여야 하도록 하고, 영향평가는 인공지능 기술 및 사업에 대하여 정확히 이해하면서도 객관성과 중립성을 확보할 수 있도록 독립된 평가팀 또는 제3의 기관이 수행하도록 하여야 할 것이다.

영향평가 결과 인권에 미치는 부정적인 영향이나 편향성, 위험성이 나타나는 경우 개선 및 방지조치를 수립하도록 하고, 개선 및 방지 조치의 이행 여부를 공개하도록 하며, 개선 및 방지 조치가 이행되기 전에는 개발과 활용을 중단하도록 권고할 수 있도록 하여야 할 것이다. 이는 인공지능 인권영향평가의 결과서를 주무기관에 제출하도록 하여 영향평가의 적절성 등을 점검하고, 관리·감독할 수 있도록 함으로써 실현 가능할 것이다. 인공지능 인권영향평가의 주무기관으로는 인권에 대한 전문성과 독립성을 확보하고 있는 국가인권위원회가 적합하다.

나. 인공지능 인권영향평가도구(안)

1) 인공지능 인권영향평가 개요

가) 인공지능 인권영향평가의 대상 : 고위험 인공지능

법률에서 금지하는 인권 침해 또는 차별 대우를 목적으로 하거나 법률에서 금지하는 개인정보의 처리를 목적으로 하는 인공지능 등 위험의 완화 내지 제거가 불가능한 인공지능은 일응 ‘금지대상 인공지능’에 해당하여 인권영향평가의 대상에서 제외된다.

인공지능 인권영향평가가 대상인 인공지능은 공공기관이 직접 개발하거나 조달하는 모든 인공지능 및 민간에서 활용하는 인권 침해의 위험성이 높은 인공지능, 즉 고위험 인공지능이다. 물론 금지되는 인공지능의 기준과 마찬가지로 현 단계에서는 무엇이 고위험 인공지능인지에 대한 명확한 사회적인 합의가 존재하지 않으며, 고위험 인공지능에 대한 인권영향평가의 실시를 의무화하는 법률도 존재하지 않는다. 고위험 인공지능에 대한 인권영향평가 실시 의무화를 위해서는 고위험 인공지능의 범위 및 인권영향평가 수행 의무화를 내용으로 하는 입법화가 선행될 필요가 있다.

그러나 인권영향평가는 인공지능의 잠재적 위험을 체계적으로 검토하고 이해관계자와의 대화와 협력을 통해 사전에 방지, 완화하자는 취지로 시행된다. 따라서 국내외에서 고위험 인공지능이라고 거론되는 인공지능의 경우 본 인권영향평가를 수행할 것이 강하게 권고된다.

- 항공, 자동차, 철도, 기계, 장난감의 안전 관련 구성요소이거나, 승강기, 무선 장비 및 의료 기기 등의 안전 관련 구성요소 또는 제품 그 자체인 경우
- 사람의 생체정보를 활용하여 신원확인을 수행하는 경우
- 교통, 수도, 가스, 전기 등 중요 사회기반시설의 관리·운용에 활용하는 경우
- 소방, 응급의료 등 필수 공공·민간 서비스에 활용하는 경우
- 채용, 인사평가 또는 직무 배치의 결정에 사용하는 경우
- 공공 지원 혜택의 자격 및 수혜 적격성을 평가하기 위하여 사용하는 경우
- 범죄의 수사, 공소의 제기 및 유지, 형 및 보안처분의 집행에 사용하는 경우

- 이주, 망명 및 출입국 관리에 활용하는 경우
- 사실의 인정 및 법률 해석, 적용 등 법관의 업무를 지원하는 데 사용하는 경우
- 군 또는 정보기관에서 사용하는 경우

나) 인공지능 인권영향평가 시기

공공기관의 경우 위험도와 무관하게, 민간의 경우 고위험 인공지능을 개발하거나, 고위험인공지능을 사업 또는 정책의 기반기술로 도입하기 이전에 인권영향평가를 수행하되, 사전영향평가에만 한정하지 않고, 정기적, 사후적 평가를 통한 지속적인 관리와 모니터링을 전제한다. 고위험인공지능에 비하여 위험의 정도가 덜한 인공지능의 경우 국가인권위원회의 직권 지정 또는 이해관계자의 요청에 따른 검토를 거쳐 국가인권위원회의 지정에 따라 인권영향평가를 수행할 수 있고, 개발 또는 도입 주체의 자발적인 요구에 의해서도 수행될 수 있다. 사전 영향평가의 시점은 인공지능기술 개발 또는 도입에 관한 구상이 구체화된 시점이다.

다) 인공지능 인권영향평가 수행 주체

구체적인 영향평가의 수행은 인공지능의 개발 주체 및 관련된 사업부서와는 독립된 별도의 조직(예를 들어, 인공지능 윤리, 인권 경영, ESG 경영 등을 담당하는 부서) 또는 독립성과 인권 분야에 대한 전문성 및 인공지능 기술에 대한 전문성을 갖춘 제3의 기관이 수행하도록 한다.

라) 인공지능 인권영향평가의 절차

본 연구는 인공지능 인권영향평가의 이행단계를 4단계로 범주화하였다.

- 1단계 : 계획 및 준비
- 2단계 : 분석 및 평가
- 3단계 : 개선 및 구제
- 4단계 : 공개 및 점검
- 공통 : 이해관계자의 참여

한편, 영향평가에 따른 결과서는 국가인권위원회에 제출되며, 국가인원위원회는 영향평가결과를 검토한 후 미흡한 점에 대한 개선을 권고하거나, 위험성에 대한 완화 조치 또는 제거 조치가 불가능하다고 판단하는 경우 개발 또는 활용의 중단을 권고하는 등 의견을 제시할 수 있도록 한다.

아래는 인권영향평가 과정에서 점검해야 할 항목을 체크리스트 방식으로 제시한다.

2) 인공지능 인권영향평가도구

【1단계 : 계획 및 준비】

가. 인권영향평가 계획

Q1-1-1. 인권영향평가의 대상이 되는 인공지능 시스템 혹은 프로젝트는 무엇입니까.
인공지능 시스템 혹은 프로젝트명 :

Q1-1-2. 인권영향평가를 수행하는 책임자는 누구입니까.
책임자의 성명과 소속 :

Q1-1-3. 평가팀은 인권영향평가를 수행하기에 충분한, 인공지능 기술 및 인권에 대한 전문성을 갖추고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
평가팀의 구성, 팀원의 역할, 전문분야 등을 설명하시오.
설명 ( )

Q1-1-4. 조직 내에 인권영향평가 수행의 요건, 주체, 절차 등을 상세히 규정한 정책을 두고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-1-5. 인권영향평가를 내실있게 수행하는데 충분한 인적, 재정적 자원이 확보되어 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-1-6. 인권영향평가 결과의 수용 여부를 결정할 수 있는 조직 내 최종 책임자 혹은 책임단위에 인권영향평가 결과보고서를 보고하는 절차가 명확하게 규정되어 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-1-7. 인권영향평가를 내실있게 수행할 수 있도록, 평가팀이 평가 대상이 되는 인공지능 시스템의 개발 혹은 활용과 관련한 부서 및 담당자에게 협조를 요청하고, 인권영향평가에 필요한 핵심 자료에 접근할 수 있는 권한이 보장되어 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-1-8. 인공지능 시스템은 어떠한 문제를 해결하기 위한 것입니까, 즉 인공지능 시스템이 달성하고자 하는 목적 및 의도된 용도는 무엇입니까.
인공지능 시스템의 목적 :

Q1-1-9. 해당 인공지능 시스템이 적용되는 분야에서 인공지능 시스템의 기능, 요건, 제한 등에 영향을 미치는, 인권 보호를 위해 요구하고 있는 법령상의 요건(법률, 시행령, 시행규칙 등의 관련 조항)이 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
만일 있다면, 해당 법률, 시행령, 시행규칙 등의 관련 조항은 무엇입니까.
설명 ( )

Q1-1-10. 인공지능 시스템이 인권에 미치는 영향을 평가하기 위하여 조직 내외부의 다양한 이해관계자의 의견을 검토할 필요가 있습니다. 다양한 이해관계자를 인권영향평가 과정에 참여시키기 위해서는 우선 누가 이해관계자인지 파악해야 합니다. 아래 질의에서 이해관계자가 누구인지 가능한 구체적으로 적어주세요.

Q1-1-10-1. 해당 인공지능 시스템에 대한 공정한 인권영향평가를 위해, 조직 내부에서

해당 인공지능 시스템의 개발 및 운영에 관련된 이해관계자(예를 들어, 기획, 개발, 디자인, 유지보수, 정책, 데이터 거버넌스, 영업 등 담당 부서)의 참여가 중요합니다. 이를 위해 조직 내부에서 참여할 수 있는 이해관계자는 누구입니까.
설명 ( )

Q1-1-10-2. 해당 인공지능 시스템에 대한 공정한 인권영향평가를 위해, 조직 외부에서 해당 인공지능 시스템의 개발 및 운영에 관련된 이해관계자(예를 들어, 외부 개발업체, 위탁업체, 유지보수업체, 감독기구, 전문가 집단 등)의 참여 역시 중요합니다. 이를 위해 조직 외부에서 참여할 수 있는 이해관계자는 누구입니까.
설명 ( )

Q1-1-10-3. 인공지능 시스템의 사용자는 누구입니까.
설명 ( )

Q1-1-10-4. 인공지능 시스템의 사용으로 영향을 받는 사람이나 집단은 누구입니까.
설명 ( )

Q1-1-10-5. 인공지능 시스템의 사용으로 영향을 받는 개인이나 집단에 아동, 노인, 장애인, 여성, 외국인, 성소수자, 저학력자, 비정규직 노동자, 경제적 약자, 낙후지역 등 취약하거나 소외된 집단이 포함되어 있다면 구체적으로 적어주세요.
설명 ( )

나. 조사

Q1-2-1. 인공지능 시스템이 인권에 미치는 영향을 이해하기 위해서는 해당 시스템에 대한 이해가 필요합니다. 데이터셋, 알고리즘 등 해당 인공지능 시스템과 관련된 정보(예를 들어, 데이터셋이나 알고리즘 등의 특성 및 이에 대한 평가, 외부업체의 제품을 구매할 경우 관련한 설명서, 사전학습 모델 가중치 등)를 확보하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-2-2. 인공지능 시스템이 도입, 활용될 분야 혹은 시공간적인 특성 및 맥락과 관련

된, 인권에 영향을 미칠 수 있는 요소에 대한 자료를 확보하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-2-3. 앞서 파악한, 인공지능 시스템의 이해관계자로부터 해당 시스템이 인권에 미칠 영향에 대한 의견을 수렴하거나 협의하고 이를 문서화 하였습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-2-4. 이해관계자 의견을 수렴하거나 협의할 때 다음과 같은 내용을 포함합니까.
- 협의한 이해관계자의 성명, 소속, 연락처
- 협의한 일자
- 인공지능 시스템에 대해 이해관계자에게 제공한 자료
- 인공지능 시스템에 대한 이해관계자의 의견
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-2-5. 인공지능 시스템의 활용으로부터 영향을 받는 이해관계자, 특히 취약하거나 소외된 집단과의 협의를 포함하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-2-6. 관련 자료를 수집하거나 이해관계자의 의견을 수렴할 때 자료의 기밀성을 유지하고 이해관계자의 개인정보를 보호할 수 있는 조치를 취하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

【2단계 : 분석 및 평가】

가. 인공지능 기술과 관련된 영향 분석 및 평가

(1) 개인정보보호

Q2-1-1. 인공지능 시스템이 개인정보보호위원회 <인공지능(AI) 개인정보보호 자율점검표>의 모든 의무/권장 조항을 준수하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-2. 해당 인공지능 시스템의 개발 혹은 운영 과정의 개인정보 처리가 개인정보 보호법 상 개인정보 영향평가를 의무적으로 수행해야 하는 경우, 개인정보 영향평가를 수행하였는지 확인하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

(2) 데이터

Q2-1-3. 학습, 검증, 테스트 등 인공지능의 개발 과정에 사용되는 데이터셋에 대한 정보, 예를 들어 데이터셋의 출처, 구조와 유형, 사전 처리 과정 등에 대한 정보를 확보하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-4. 데이터셋의 정확성, 완전성, 최신성을 확인하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

이를 검토하기 위해 사용한 방법은 무엇입니까.

설명 ( )

Q2-1-5. 데이터셋이 인공지능 시스템이 사용될 맥락에 적합하도록 인구집단별 다양성과 대표성을 갖추었는지 확인하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

이를 검토하기 위해 사용한 방법은 무엇입니까.

설명 ( )

Q2-1-6. 데이터셋이 사상·신념, 건강, 인종이나 민족에 관한 정보, 생체인식정보 등 민감정보를 포함하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-7. 대리 변수를 통해 민감정보의 추정이 가능한지 여부를 검토하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-8. Q2-1-3 ~ Q2-1-7의 질의 전체 혹은 일부에 대한 확인이 불가능하거나 이를 확인하는 것이 불필요하다고 판단하는 경우, 그 이유는 무엇입니까. 또한 그러한 경우 데이터셋의 편향성을 방지할 수 있는 다른 방안은 무엇입니까.

설명 ( )

(3) 알고리즘의 성능과 신뢰성

Q2-1-9. 인공지능 시스템 외의 다른 대안 혹은 채택된 알고리즘(또는 사전학습 모델 가중치) 외에 다른 대안에 대한 검토가 있었습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

인공지능 시스템에 사용된 알고리즘(또는 사전학습 모델 가중치)이 목적 달성에 적합한 이유는 무엇입니까.

설명 ( )

Q2-1-10. 인공지능 시스템이 의도한대로 작동하는지 성능을 측정하기 위한 지표와 방법을 갖고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-11. 정확도와 오류율 등 성능의 수준은 의도한 목적에 적합한 정도로 설정되었습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

인공지능 시스템의 정확도와 오류율 등 성능은 어떻게 측정합니까.
설명 ( )

(4) 차별금지

Q2-1-12. 인공지능 시스템이 활용 과정에서 합리적인 이유없이 인종, 종교, 장애, 나이, 학력, 직업, 출신 지역, 언어, 정치성향, 신체조건, 외모, 피부색, 병력, 성별, 성적 지향, 사회적 신분, 경제적 지위 등 개인과 집단의 특성에 따라 특정 집단에 대한 차별을 야기하거나 혹은 기존의 차별을 악화할 가능성이 있는지 검토하였습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q2-1-13. 인공지능 시스템 개발 과정에서 알고리즘에 의한 구조적 차별을 사전에 방지하기 위하여, 기획, 개발, 디자인, 마케팅, 경영진 등 조직 구성원의 다양성 확보, 구성원에 대한 반차별 교육, 조직 내 인공지능 윤리 정책 수립 등의 대책을 마련하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

(5) 설명가능성과 투명성

Q2-1-14. 해당 인공지능 시스템이 특정한 결정(출력)을 내리는데 관련된 요소를 추적할 수 있도록 관련된 정보(예를 들어, 결정의 내역이나 시스템에 대한 모든 변경사항 등에 대한 로그기록)가 기록되고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q2-1-15. 해당 인공지능 시스템이 특정한 결정(출력)을 내린 이유나 근거에 대해 사용자 혹은 영향을 받는 이해관계자에게 설명할 수 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q2-1-16. 설명가능성 여부가 인권에 영향을 미칠 경우, 해당 인공지능 시스템의 작동이나 특정한 결정의 근거에 대해 기술 전문가가 아닌 이해관계자가 충분히 이해할 수 있는 방식으로 설명할 수 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-17. 정확도와 오류율 등 인공지능 시스템의 성능, 어떤 결정을 내리는데 사용되는 매개변수 및 가중치, 적절한 사용법, 장점과 한계 등에 대해 사용자가 이해할 수 있는 방식으로 충분한 정보를 제공하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-18. 인공지능 시스템의 소스코드가 공개되거나 이를 요구하는 이해관계자에게 제공될 수 있습니까. 소스코드가 제공될 수 있다면, 누구에게 어떤 조건으로 제공됩니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

(6) 자동화 정도와 인간의 개입

Q2-1-19. 사람과 상호작용하는 인공지능 시스템의 경우, 인공지능 시스템의 사용자 혹은 상호작용하는 사람에게 상대방이 사람이 아니라 인공지능 시스템이라는 사실, 혹은 자신이 받은 결과물이나 결정이 인공지능 시스템에 의한 것이라는 점을 적절하게 알릴 수 있는 조치를 취하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-20. 인공지능 시스템이 영향을 받는 이해관계자가 인지할 수 없도록 은밀하게 작동할 수 있는 경우(예를 들어, 원격에서 작동하는 얼굴인식 시스템이 대상자 모르게 얼굴인식을 통해 신원을 파악하는 경우) 영향을 받는 당사자가 인지하지 못하는 방식으로 인공지능 시스템이 작동하지 않도록 당사자에게 적절하게 알릴 수 있는 조치를

취하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-21. 인공지능 시스템의 결과물에 기반한 결정에서 인간의 역할과 인간이 재량권을 갖고 개입할 수 있는 범위와 절차가 정의되어 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-22. 인공지능 시스템이 의도한대로 작동하지 않을 경우, 인공지능 시스템의 운영자 혹은 사용자는 언제든지 시스템을 정지시킬 수 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

(7) 보안

Q2-1-23. 인공지능 시스템의 특성과 활용되는 분야 등을 고려했을 때, 인공지능 시스템 보안에 대한 가능한 위협이 무엇이고 보안이 침해되었을 경우 발생할 수 있는 결과 혹은 해악은 무엇입니까.

설명 ( )

Q2-1-24. 인공지능 시스템의 학습 및 테스트에 활용되는 데이터셋에 대해 충분한 안전조치가 적용되었습니까. 데이터 오염 등 데이터에 대한 다양한 유형의 공격에 대한 대응이 고려되었습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-25. 인공지능 시스템의 전체 수명주기 동안 발생할 수 있는 잠재적인 공격에 대비하여 무결성, 가용성, 기밀성, 견고성 등 보안에 요구되는 요소를 보장하기 위한 조치를 취했습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

(8) 접근성

Q2-1-26. 인공지능 시스템이 언어, 나이, 장애, 신체적 조건 등에 상관없이 누구나 사용할 수 있도록 인공지능 시스템의 인터페이스가 보편적 설계 원칙에 따라 설계되었습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-27. 보편적 설계 원칙에 따라 설계하지 않은 합리적인 이유가 있다면 그것은 무엇입니까.

설명 ( )

(9) 라이선스

Q2-1-28. 인공지능 시스템의 전체 혹은 일부 소프트웨어를 외부에서 개발된 것을 사용할 경우, 인공지능 시스템에 의한 잠재적 인권 침해를 방지, 완화하기 위하여 필요한 경우 알고리즘 혹은 소스코드 등 소프트웨어를 적절하게 수정, 변경할 수 있는 권한에 대해 외부 개발업체와 명확한 합의가 이루어져 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

나. 인권에 미치는 영향 및 심각도

(1) 영향을 받는 인권

Q2-2-1. 인공지능 시스템이 도입, 활용될 경우 시민들의 인권에 미칠 수 있는 부정적인 영향 혹은 위험은 무엇입니까. 누구의 인권이 어떤 방식으로 침해될 수 있습니까.

설명 ( )

Q2-2-2. 인공지능 시스템이 오류로 인하여 의도하지 않은 방식으로 작동할 경우 나타날 수 있는 부정적인 결과에 대해 검토한 바 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
의도하지 않은 방식으로 작동할 경우 침해되는 인권은 무엇입니까
설명 ( )

Q2-2-3. 인공지능 시스템이 의도적으로 악용될 가능성이 있습니까. 어떠한 방식으로 오용될 수 있는지에 대해 검토하였습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
악용될 경우 나타날 수 있는 부정적인 결과, 혹은 침해되는 인권은 무엇입니까
설명 ( )

(2) 인권에 미치는 영향의 심각도

Q2-2-4. 인공지능 시스템이 인권에 미치는 부정적 영향의 범위가 어떠합니까. 전체 인구 혹은 어떠한 특정 집단에 대하여 어느 정도의 범위(대, 중, 소)로 영향을 미칠 수 있습니까. (부정적 영향을 받을 수 있는 인권이 여러 개인 경우 각각에 대해서 평가가 필요함. 아래 질의에 대해서도 동일함)
설명 ( )

Q2-2-5. 인공지능 시스템이 인간의 생명, 건강, 안전, 인권, 기본적 삶 등에 미치는 부정적 영향의 규모 혹은 크기가 어떠합니까. (대, 중, 소)
설명 ( )

Q2-2-6. 인공지능 시스템이 인권에 미치는 부정적 영향이 사후에 구제나 회복이 어느 정도 가능합니까. (완전히 회복 가능, 부분적으로 회복 가능, 회복 불가능)
설명 ( )

Q2-2-7. 인공지능 시스템이 인권에 미치는 부정적 영향이 여러 개일 경우 서로 상충하는 인권이 있습니까. 또한 그 심각성 때문에 우선적인 대응이 필요한 인권은 무엇입니까.
설명 ( )

【3단계 : 개선 및 구제】

가. 방지

Q3-1-1. 데이터에 대한 개선, 알고리즘의 수정, 시스템 설계 변경 등 2단계(영향 분석 및 평가)에서 파악된 중대한 인권 위험을 방지하기 위해 어떠한 조치를 취했습니까.
설명 ( )

나. 완화

Q3-2-1. 2단계(분석 및 평가)에서 파악된 인권 위험을 완전히 방지하기 곤란한 경우, 위험을 완화하기 위해 어떠한 조치를 취했습니까.
설명 ( )

Q3-2-2. 인공지능 시스템의 잔존하는 위험성에 대해 사용자 및 영향을 받는 이해관계자에게 충분한 정보를 제공하고 올바른 작동 방법에 대해 적절한 교육을 제공하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q3-2-3. 인공지능 시스템의 인권 침해 위험이 클 수 있는 특정한 사용을 허용하지 않도록 이용약관이나 여타의 집행체계에서 금지하는 절차를 취하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q3-2-4. 2단계(분석 및 평가)에서 파악된 중대한 인권 위험이 완화되지 않고 남아있을 경우 그 이유를 문서화하고 있습니까
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q3-2-5. 중대한 위험에 대한 방지 및 완화 조치를 취하기 힘들거나, 이러한 조치를 취해도 여전히 중대한 위험이 남아있을 경우 인공지능 시스템의 개발 및 활용을 중단할

계획을 갖고 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

다. 구제

Q3-3-1. 인공지능 시스템의 결정에 의해 영향을 받는 사람이 인공지능 시스템의 결정에 이의를 제기하거나 침해된 권리의 구제를 요구할 수 있는 절차를 마련하고, 이에 대한 정보를 누구나 쉽게 접근할 수 있도록 일반에 공개하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q3-3-2. 인공지능 시스템에 의해 영향을 받는 사람들에게 인공지능 시스템의 사용을 거부할 수 있는 선택권(옵트아웃 권리)을 제공하거나 이의를 제기할 수 있는 수단이 마련되어 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q3-3-3. 인공지능 시스템이 내린 결정에 의해 영향을 받는 이해관계자가 인공지능 시스템의 적용을 거부할 경우, 인간의 지원 혹은 인공지능이 아닌 시스템의 적용을 대안으로 제시할 것을 고려하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q3-3-4. 인공지능 시스템의 결정에 대한 이의제기나 권리구제 요구가 정당할 경우, 문제의 의사결정을 번복하거나 권리를 복구하거나 손해배상을 할 수 있는 절차가 마련되어 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

라. 이해관계자와의 의견수렴 및 협의

Q3-4-1. 인공지능 시스템의 인권 위험을 방지, 완화하고 인권을 침해받은 사람의 권리를 구제하기 위한 조치에 대해서 관련 이해관계자(1단계에서 파악한 이해관계자)의 의견을 수렴하거나 협의를 진행하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q3-4-2. 공공기관이 도입하는 인공지능 시스템의 경우, 가능한 모든 이해관계자가 참여할 수 있도록 의견 수렴을 위한 공청회를 실시하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

【4단계 : 공개 및 점검】

가. 인공지능 시스템의 주요 요소의 공개

Q4-1-1. 인공지능 시스템이 사용하는 데이터와 알고리즘 등의 주요 요소를 일반에 공개하고 이해할 수 있는 방식으로 쉽게 설명하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

나. 인권영향평가 결과 공개

Q4-2-1. 인권영향평가 보고서 전체 혹은 주요 내용을 일반에 공개합니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q4-2-2. 인권영향평가 보고서를 감독기구인 국가인권위원회에 제공하고, 효과와 한계에 대해 협의하는 절차가 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

다. 인공지능 시스템에 대한 모니터링

Q4-3-1. 인공지능 시스템이 도입되거나 운영이 시작된 후에 그 성능과 인권에 미치는 부정적 영향, 완화 조치 및 구제 정책의 효과성을 확인하기 위해, 인공지능 시스템의 수행을 모니터링하고 기록에 남길 수 있는 메커니즘을 수립하였습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q4-3-2. 인공지능 시스템이 의도한 대로 작동하지 않거나 인권에 미치는 부정적인 영향이 확인되었을 때, 관련된 책임을 명확히 하고 인공지능 시스템을 개선하며 부정적 영향을 완화하기 위해 필요한 절차를 수립하였습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

라. 인권영향평가에 대한 점검

Q4-4-1. 인권영향평가 수행의 효과와 한계를 점검하고, 개선할 수 있는 절차를 마련하였습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

마. 인권영향평가의 재수행

Q4-5-1. 인공지능 시스템에 대해 정기적으로 (예를 들어 1년) 인권영향평가를 수행하는 절차를 마련하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q4-5-2. 인공지능 시스템의 핵심적인 기능이 변경되거나 환경적 요인 혹은 적용 범위가 변경되었을 경우, 인권영향평가를 다시 수행하는 절차를 마련하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

한편, 인공지능 인권영향평가 1단계에서 평가팀은 여러 이해관계자로부터 자료와 의견을 수집하게 되는데, 인공지능 시스템 자체에 대한 정보를 파악하기 위해서는 해당 시스템을 기획, 개발, 설계, 디자인한 사람이나 팀으로부터 관련 자료를 얻을 수밖에 없다. 외부의 인공지능 기술, 시스템, API 등을 활용할 경우에는 해당 업체로부터 관련 자료를 얻게 될 것이다. 평가팀은 인권영향평가도구의 질의에 답을 하기 위해 개발팀 혹은 개발업체에 자료를 요청하거나 의견을 구할 수 있다. 평가팀은 개발자가 제공한 정보나 답변을 다른 이해관계자가 제공한 정보나 답변, 그리고 평가팀이 자체적으로 평가한 내용과 비교, 분석함으로써 객관적인 평가를 수행할 수 있다.

제1장 서론

제1절 연구 목적 및 필요성

최근 몇 년간 국가와 기업에서 인공지능을 개발하고 활용하는 사례가 급격히 증가하고 있다. 인공지능의 발전은 분석과 예측 등 디지털 정보의 활용 가능성을 증대시켜 업무와 시스템의 효율성을 크게 향상시켰고, 중요한 의사결정을 지원하거나 직접 자동화된 결정을 수행하는 수준에 이르렀다. 그러나 인공지능이 인권에 미치는 부정적인 영향과 위험에 대한 우려도 그만큼 커지고 있다. 인공지능 챗봇 이루다는 소수자를 혐오하는 내용의 채팅으로 우리 사회에 큰 충격을 주었다.[1] 코로나19 위기를 거치면서 수많은 공공기관과 기업이 비대면으로 인공지능 채용 절차를 실시하였으나, 불투명하고 공정하지 않다는 의혹을 받고 있다.[2]

「지능정보화 기본법」에서 인공지능은 주로 "전자적 방법으로 학습·추론·판단 등을 구현하는 기술"을 의미한다(제2조 제4호 가목). 유럽평의회는 알고리즘 시스템의 인권영향에 대한 권고 CM/Rec(2020)1에서 이러한 정의를 좀더 구체화하여 '알고리즘 시스템'이란 "종종 수학적 최적화 기술을 이용하여 데이터 수집, 결합, 정리, 정렬, 분류 및 추론 뿐 아니라 선택, 우선 순위, 권장 사항 및 의사결정과 같은 업무를 하나 이상 수행하는 응용프로그램"이라고 정의하며 "적용되는 설정에서 요구사항을 충족하기 위해 하나 이상의 알고리즘에 의존하는 알고리즘 시스템은 대규모 및 실시간으로 적응형 서비스를 생성하는 방식으로 업무를 자동화한다"고 설명한다. 즉, 최근 논의되고 사용되는 인공지능은 자의식을 가지고 있다거나 범용으로 쓸 수 있는 '강(強)인공지능'이 아니라, 인간이 요구하는 특정 목표에 따라 학습하고 추론하고 판단하는 방식으로 데이터와 업무를 처리하는 컴퓨터 응용프로그램이다.

과거보다 부쩍 스마트해진 컴퓨터 응용프로그램으로서 인공지능은 일반적으로 대규모 데이터셋에서 패턴을 감지하여 작동한다. 특히 정밀도, 타겟팅 및 일관성 면에서 업무 및 시스템 성능의 효율성 및 효과성을 크게 증진시켰고 디지털 정보의 분류 및 검색 기

1) 연합뉴스(2021. 1. 11). 성희롱·혐오논란에 3주만에 멈춘 '이루다'…AI윤리 숙제 남기다.
2) 한겨레21(2020. 10. 23). AI 면접관이 말했다 "너 인성 문제 있어?".

능을 크게 향상시켜 의료 진단, 운송 및 물류와 같은 분야에서 성취를 달성해 왔다. 특히 인공지능의 분석과 예측 기능은 부분적으로나 완전히 자동화된 방식으로 인간의 의사결정을 지원하거나 대체하는 수준에 이르러 행정 처분과 재판 업무에 쓰이기도 한다. 이러한 배경에 힘입어 2021년 제정·시행된 행정기본법(제20조)이나 개정·시행된 전자정부법(제18조의2)은 행정기관이 인공지능 기술을 활용하거나, 완전히 자동화된 방식으로 행정 처분을 내리는 것을 가능케 하였다.

그러나 인공지능이 개인정보를 이용하여 학습하고 사람들의 삶에 영향을 미치는 결정을 내리게 되면서 개인정보 및 사생활의 권리가 큰 영향을 받게 되었다. 개인정보를 이용하지 않는 인공지능이라 하더라도 그 추론과 예측은 사생활의 권리에 깊은 영향을 미치며, 인공지능이 편향적으로 개발되고 활용될 경우 차별받지 않을 권리에도 부정적인 영향을 미칠 수 있다. 또한 인공지능이 적용되는 영역별로 건강권, 교육권, 이동의 자유, 평화적인 집회의 자유, 결사의 자유, 표현의 자유 등 여타 인권의 향유도 다양한 영향을 받는다.

유엔 인권최고대표는 결함이 있거나 편향적인 데이터에 기반하여 인공지능 시스템의 결과물이 산출될 경우, 한 개인을 테러범이나 부정수급자로 잘못 지목함으로써 인권 침해를 야기할 수 있다고 우려하였다. 유럽평의회는 이러한 인권 침해 우려가 합리화와 정확성의 향상으로만 상쇄되기는 어렵다고 지적한다. 대부분의 알고리즘 시스템은 오류가 불가피한 통계 모델을 기반으로 하며, 기존의 편향, 오류 및 가정을 유지, 복제 및 강화하는 피드백 순환구조를 가지고 있다는 사실에 주목할 필요가 있다는 것이다. 알고리즘 시스템이 많은 사람들에게 사용되면 될수록, 위양성 및 위음성 등의 오류와 내재된 편향의 영향을 받는 사람들의 수도 증가하여 다양하고 추가적인 인권 침해가 유발될 위험 역시 커진다.

따라서 이와 같이 인공지능이 인권에 미치는 부정적인 영향과 위험을 방지하고 완화하기 위해서는 인공지능이 배치되거나 사용되기 전에 예방적 조치를 취하는 것이 반드시 필요하다. 최근 세계 각국에서 다양한 인공지능 영향평가를 검토하고 도입하는 이유 역시 바로 이와 같은 맥락에 있을 것이다. 유럽연합, 캐나다, 영국, 미국에서는 인공지능에 대한 다양한 영향평가 제도를 법제도적 수준으로 이미 도입하였거나 제안하는 절차를 밟고 있다. 한편 유엔 인권기구, 유럽평의회, 덴마크, 네덜란드 등에서는 인공지능에 대한

인권 기반 접근법(rights-based approach)을 취하며 인권영향평가를 제안해 왔다. 기존 제도로 관리하거나 감독할 수 없는 새로운 분야에서는 인권영향평가의 필요성이 더욱 크다고 할 수 있다.

그러나 현재까지 인공지능에 대한 영향평가, 특히 인권영향평가의 내용과 제도화에 대한 연구가 많이 이루어지고 있지 못한 실정이다. 본 연구는 국내외에서 인공지능에 적용하였거나 준비 중인 다양한 영향평가 제도를 우선 발굴하여 다각도로 검토한 후에, 특히 도입 가능한 인공지능 인권영향평가도구와 그 실현을 위한 제도적 방안을 정책적으로 제언하고자 한다.

제2절 연구 내용 및 범위

본 연구는 다음과 같은 연구 범위를 가지고 있다. 첫째, 인공지능 인권영향평가에 대한 구체적인 사례를 검토한다. 특히 영역별, 단계별로 다양한 영향평가 실행에 대한 국제기준 및 해외 정책사례를 검토하고 이어서 인공지능 인권영향평가도구에 대한 국제인권기준 및 해외 정책사례를 검토한다. 둘째, 검토 결과를 토대로 인공지능 인권영향평가에 대한 구체적인 정책을 제언한다. 여기서는 규범력을 갖춘 제도화 방안 및 인권영향평가의 적용 기준을 포함하고, 전문가 심층면접조사를 통하여 개선된 제언을 도출한다. 마지막으로 해외사례와 비교 분석을 실시한다. 인권영향평가를 비롯하여 해외에서 제안되었거나 도입 중인 다양한 인공지능 영향평가의 기준 및 정책사례를 본 연구 결과 도출된 제도화 방안 및 적용 기준과 함께 비교 분석하여 그 타당성을 살펴 본다.

이에 우선 2장에서는 인권영향평가에 대한 선행 검토를 수행한다. 인권영향평가의 기준 및 실행 사례 검토를 통하여 인공지능 인권영향평가에 대한 시사점을 도출한다. 이어 3장에서는 인공지능 영향평가의 사례를 검토한다. 우선 유럽연합, 캐나다, 영국, 미국 등 위험영향평가를 비롯하여 인공지능에 대하여 다양하게 도입 중인 영향평가제도를 살펴본다. 이어서 유엔 인권규범, 유럽평의회, 덴마크 네덜란드에서 제안하였거나 도입한 인공지능 인권영향평가제도의 기준을 검토하고 주요 빅테크 기업의 실행 사례를 살펴본다. 더불어 국내적으로 각 부처와 지방자치단체에서 제안해 온 인공지능 자율점검 기준을 일별한 후 국가인권위원회 인권영향평가 정책을 중심으로 적용 가능성을 살펴본다.

4장에서는 인권영향평가 절차 및 도구에 대한 초안을 개발하여 전문가에 대한 심층면접조사를 실시하고, 5장은 이상에서 검토한 내용을 종합하여 인공지능 인권영향평가의 제도적 측면과 기준을 제언한다. 6장은 연구진의 최종적인 결론 및 정책 제언으로 마무리한다. 연구 과정에서 산출된 번역 결과물은 부록으로 담는다.

제3절 연구 방법

본 연구는 우선 국내·외 문헌에 대한 연구를 통하여 인권영향평가를 비롯한 인공지능 영향평가 제도에 대한 보고서 등 선행연구와 관련 행정자료를 조사하였다. 세계 각국 및 국제기구의 관련 규범 및 법률(안) 등 주요 사례에 대한 검토도 수행하였다.

특히 유럽연합, 캐나다, 영국, 미국 등에서 도입 중인 다양한 인공지능 영향평가와 더불어 유엔 인권규범, 유럽평의회, 덴마크, 네덜란드에서 제안하였거나 도입 중인 인공지능 인권영향평가제도의 기준을 조사하고 주요 빅테크 기업의 실시 사례를 검토하였다.

본 연구는 이러한 검토 결과를 토대로 인공지능 인권영향평가(안)을 개발하였다. 우선 초안을 개발하여 기술자, 법률가, 기업, 학술연구자, 공공기관은 물론 영향을 받는 당사자로서 여성, 수급대상자, 장애인, 이주민, 학교구성원, 지역주민이나 이를 대표하는 인권시민단체 활동가의 상세한 검토를 거치는 개별서면조사를 실시하였다. 이러한 의견수렴 결과를 거쳐 최종적인 인공지능 인권영향평가도구(안)을 마련하였다.

제2장 인권영향평가 제도 연구

인권영향평가(Human Rights Impact Assessment, HRIA)는 사업과정, 정책, 입법, 프로젝트 등이 인권에 미치는 영향을 측정하고 평가하는 도구이다. 즉, 인권영향평가는 국가, 지방자치단체, 기업 등 공적·사적 주체들이 시행·추진하는 사업과정이나 정책 등에서 인권에 미치는 부정적 영향을 방지·완화하고, 긍정적인 영향을 미치는 행위를 하도록 장려하기 위해, 법령과 정책 및 사업 등의 계획과 활동이 인권의 실현과 보호에 부합하는지를 평가하고 검토하는 것을 말한다.[3)]

일반적으로 인권영향평가는 평가시점에 따라 사전적·사후적 인권영향평가로 유형화할 수 있지만, 근거규범이나 실시유무, 결과반영의무에 따라 구분하기도 하고, 주도하는 주체에 따라 기업 주도의 인권영향평가와 이해관계자 주도의 인권영향평가로 구분하기도 한다.[4)]

인공지능 인권영향평가도 넓은 의미에서 인권영향평가에 속하는 도구이므로, 인공지능 인권영향평가 도입에서 국내외의 전반적인 인권영향평가를 살펴봄으로써 일정한 시사점을 얻을 수 있다. 여기에서는 국내외 인권영향평가의 내용을 개략적으로 살펴보고 인공지능 인권영향평가 도입에서 얻을 수 있는 시사점을 도출하고자 한다.

제1절 국내외 인권영향평가 현황 및 사례

1. 해외 인권영향평가 현황 및 사례

유엔은 인권영향평가의 원칙과 지침을 형성하는 데 주된 역할을 했다. 특히 유엔 인권이사회(UN Human Rights Council)의 두 가지 지침이 표준 문서로 많이 거론된다.[5)] 2011

3) 인권영향평가의 개념에 대하여는 정영선(2013), 110면; 최유(2015), 430면; 이충은·노진석(2018), 219면 등 참조. 우리나라에서 인권영향평가의 개념에 대해서는 아직 학문적으로나 제도적으로 정립되지 않았다. 다만 국내의 인권영향평가에 관한 연구나 문헌에서 평가의 대상을 국가나 지방자치단체와 같은 공기관에 한정하는 경향이 있으나, 국제사회는 인권영향평가의 대상으로 기업의 사업 또는 정책도 포함하는 것이 일반적이라는 점에 주의하여야 한다. 이에 대하여는 박준석(2022), 2면 참조.

4) 김종철 외(2020), 18면 이하 참조.

년의 「기업과 인권에 관한 이행지침(UN Guiding Principles on Business and Human Rights, UNGPs, 이하 '기업과 인권 이행지침')」과 2018년의 「경제개혁 인권영향평가 지침(UN Guiding Principles on Human Rights Impact Assessments of Economic Reforms)」이 그것이다.

가. 2011년 「기업과 인권에 관한 이행지침」

이행지침은 2011년 유엔 사무총장의 특별대표였던 존 러기의 주도하에 제안되었고, 같은 해 7월에 유엔 인권이사회가 만장일치로 승인하였으며, OECD, 각국 국가인권기구, 국제표준화기구(ISO), 기업, 금융기관, NGO 등에 의해 광범위한 지지를 받고 있어 인권영향평가에 관한 국제표준으로 인식되고 있다.[6)]

이 이행지침은 국가의 인권보호의무, 기업의 인권존중책임, 효과적인 구제수단에 대한 접근이라는 세 가지 프레임워크의 실행을 제안했다(보호・존중・구제 프레임워크). 특히 기업경영과 관련하여, 기업은 국제적으로 승인된 모든 인권을 존중해야 원칙을 천명했다. 이를 위해 기업의 인권존중 정책서약, 인권실사(Human Rights Due Diligence, HRDD) 실시, 인권침해에 대한 구제조치를 요구했다.[7)] 또한 국가의 인권보호의무와 기업의 인권존중책임을 분리하여 선진국이든 개도국이든 이에 진출한 기업은 해당 국가의 인권보호법제와 독립적으로 인권존중책임을 부담하게 한 데에도 의의를 찾을 수 있다.

인권영향평가와 관련하여 주목되는 부분은 인권실사이다.[8)] 이 개념은 이행지침이 천명하는 기업의 인권존중책임의 핵심이라 할 수 있다. 이행지침 제17문은 인권실사의 내용을 다음과 같이 기술하고 있다.[9)]

---

5) 박준석(2022), 4면 이하; 이상수(2015), 72면 이하.

6) 이상수(2015), 72면.

7) 위의 글.

8) 국내에서 인권실사에 해당하는 human rights due diligence는 '인권에 대한 상세한 주의', '인권 상세주의 의무', '실천점검의무' 등 다소 혼란스럽게 번역된 바 있다. 여기에서는 "인권실사"로 표기한다.

9) 이하 번역은 이상수(2015), 73면에서 인용하였다. 이행원칙의 전체 번역은 국가인권위원회(2011)을 참조할 수 있는데, 여기에서는 인권실사를 '인권에 대한 상세한 주의'로 옮겼다.

17. 부정적 인권영향을 식별하고 방지하고 완화하며 어떻게 그에 대처하는지를 설명하기 위해서 기업은 인권실사를 수행해야 한다 이 절차는 실제적, 잠재적 인권영향을 평가하는 것, 그 결과를 [경영에] 통합하고 실행하는 것, 반응을 추적하는 것, 영향에 어떻게 대처하는지에 대해 소통하는 것을 포함해야 한다. 인권실사는,

(a) 기업이 자신의 활동을 통해서 유발하거나 기여할 수 있는 부정적 인권영향 또는 그 사업관계에 의해 자신의 사업활동, 제품 및 서비스와 직접 연결될 수 있는 부정적 인권영향을 모두 다루어야 한다.

(b) 기업의 크기, 중대한 인권영향의 위험, 그리고 사업활동의 성격과 맥락에 따라 달라야 한다.

(c) 계속적인 것이어야 한다. 이는 기업의 사업활동과 사업맥락이 전개되면서 시간이 감에 따라 인권위험이 변화할 수 있다는 것을 인정해야 한다는 것이다.

인권영향평가는 인권실사의 핵심도구이다. 인권실사는 ① 인권영향평가의 실시(identify), ② 내부통합(integration), ③ 추적 및 검증(verify), ④ 소통(communication)의 네 가지 과정으로 이루어져 있는데, 그 중심축은 역시 인권영향평가라 할 수 있다.[10)]

첫째, 기업은 인권영향평가를 실시함으로써 기업의 자체적 활동이나 사업관계에서 발생한 모든 실제적・잠재적인 부정적 인권영향을 식별하고 평가한다.

둘째, 인권영향평가의 결과를 기업내 관련 기능과 절차 전체에 통합하고, 적절한 조치를 취해야 한다. 여기에서는 내부 의사결정, 예산 분배, 감시 절차 등이 포함된다.

셋째, 양적・질적 지표에 기초하여 이해관계자의 반응을 추적해 인권영향에 대한 기업의 조치가 적절했는지 검증한다.

넷째, 인권영향평가에 관한 이상의 절차와 결과를 외부에 설명하고 관련 정보를 공개한다.

이상의 인권실사는 계속적인 과정으로 기업의 사업활동과 사업맥락에 따라 인권 위험은 달라질 수 있다는 점도 지적하고 있다.

10) 이상수(2015), 74면.

나. 2018년 「경제개혁 인권영향평가 지침」

2018년 발표된 「경제개혁 인권영향평가 지침」은 유엔 독립전문가의 보고서로 작성·발표되고 2019년 3월 유엔 인권이사회에서 채택되었다. 이 지침은 인권에 부합하는 경제개혁 정책의 기본적 조건을 제시하고 있는데, 공공정책에 대한 인권영향평가의 그간의 논의를 종합한 실천 규범이라는 평가를 받고 있다.[11)]

이 지침은 22가지 원칙을 5개로 분류하여 제시하고 있는데, ① 경제정책과 인권에 관한 국가 또는 지방정부의 의무, ② 적용가능한 인권 기준, ③ 정책의 구체화, ④ 국가, 국제금융기관, 사적 주체의 기타 의무 그리고 ⑤ 인권영향평가를 내용으로 한다.[12)]

특별히 지침은 제17조부터 제22조까지 인권영향평가에 관한 원칙을 제시하고 있다.

> 17. 국가는 경제위기 시기나 정상적인 시기 모두 인권영향평가를 시행하여야 한다.
>
> 18. 이러한 인권영향평가의 목적은 제안된 정책의 장단기, 중기의 인권영향을 평가하는 것이어야 한다. 이를 위해 국가는 정책 채택 전에 인권에 대한 잠재적 영향을 평가할 수 있도록 인권영향평가를 실시해야 한다.
>
> 19-20. 인권영향평가를 수행하는 과정에서 참여, 정보접근, 책임의 원칙을 준수하여야 한다. 소외집단 및 특별히 위험에 처한 집단을 포함한 모든 사람들의 효과적이고 시의적절하며 의미있는 참여를 통해 최대한 광범위한 전국적 소통(national dialogue)이 가능할 수 있도록 하고 이를 추구하는 것이 중요하다. 공공 재정의 모든 측면에 대한 포괄적이고 접근가능한 정보가 적시에 제공되는 경우에만 진정한 참여가 가능하다.
>
> 21. 경제개혁 정책의 설계 그리고(또는) 시행에서 사법에 대한 접근권 및 정책의 실행 및 누락에 따른 효과적 구제에 대한 권리가 보장되어야 한다.
>
> 22. 인권영향평가는 해당 국가의 적용가능한 기준을 준수하고 젠더에 대한 고려사항을 존중하는, 독립적이고 신뢰할 수 있는 인권영향평가를 생산할 수 있는 가장 적합한 기관이 책임을 맡아야 한다.

11) 박준석(2022), 4-5면 참조.
12) United Nations Human Rights Special Procedures(2020), pp.5-9.

이 지침은 국가의 인권영향평가 시행 의무, 인권영향평가의 목적, 평가의 시기, 준수원칙, 평가의 주체 등에 관한 지침을 제공하고 있다.

다. 2010년 「인권영향평가 및 관리에 관한 지침」

2010년의 「인권영향평가 및 관리에 관한 지침(Guide to Human Rights Impact Assessment and Management, HRIAM)」은 앞서 두 지침보다 앞서 발표된 지침으로 국제기업 지도자 포럼(International Business Leaders Forum, IBLF)과 국제금융공사(International Finance Corporation, IFC) 그리고 유엔 세계협약(United Nations Global Compact, UNGC)의 공동 작업이었다.[13] 이 지침은 기업 활동으로 초래되는 인권 위험과 영향을 평가하기 위한 인권영향평가의 원칙과 절차를 단계별로 비교적 상세히 제시하고 있다.

이 지침에서 인권영향평가의 기준은 세계인권선언, 시민적 및 정치적 권리에 관한 국제규약(ICCPR), 경제적, 사회적 및 문화적 권리에 관한 국제규약(ICESCR)을 기본으로 하고 있다. 특히 인권영향평가의 절차를 다음과 같은 7단계로 제시한다.[14]

1. 준비(Preparation)
   - 자사의 인권실사 접근방법 결정하기
   - 자사의 인권영향평가의 범위 정하기
2. 확인(Identification)
   - 주요 인권 위험과 영향 확인하기
   - 기준 설정하기
3. 참여(Engagement)
   - 인권 위험과 영향의 검증에 이해관계자 참여시키기
   - 인권 이슈를 고려하는 고충처리 메커니즘 개발하기
4. 사전평가(Assessment)

---

13) 이 지침의 번역으로 국가인권위원회(2014) 참조.
14) 아래 절차는 이준일 외(2018), 57면을 수정하여 인용.

- 인권 위험과 영향을 사전평가 한다.
- 사전평가 결과를 분석한다.

5. 완화(Mitigation)
   - 적절한 완화 행동 계획을 개발한다.
   - 완화 행동 계획과 권고를 관리부서에 알린다.
6. 관리(Management)
   - 완화 행동 계획과 권고를 이행한다.
   - 인권을 관리 시스템에 통합시킨다.
7. 사후 점검(Evaluation)
   - 인권을 다루는 기업 능력을 감시, 평가, 보고하기
   - 사후평가를 검토하고 필요한 경우 적절히 수정하기

이 지침은 12개 분야에 대한 가상의 인권 관련 사례를 소개하여 인권영향평가에 활용하도록 하고 있다. 특히 앞서 이행지침상 기업의 인권존중책임에서 기업이 존중해야 하는 '인권'을 35개의 목록으로 제시하고 있다는 점에서 특징적이다.[15]

라. 2020년 덴마크 인권기구의 「인권영향평가 가이드 및 도구모음」

해외의 인권영향평가 관련 가이드라인은 대체적으로 주요 절차와 절차별 요구사항 또는 주의사항을 제시하고 있다.[16] 최근 들어 구체적인 체크리스트나 템플릿도 제안되고 있어 구체적인 인권실무에 참고할만하다.[17]

특히 인권영향평가 분야에서 최근 가장 주목받고 있는 예는 덴마크이다. 덴마크 국가인권기구(The Danish Institute for Human Rights)는 기업활동, 교육, 환경, 보건 등 다양한 분야의 인권영향평가 지침과 실행도구를 개발하여 발표해 왔으며, 공적·사적 활동들과 '인권'을 통합하기 위한 지표(indicator)를 정립하기 위한 연구를 진행하고 있다. 덴

15) 국가인권위원회(2014), 99면 참조.
16) 한국도시연구소(2019), 4면.
17) 대표적으로 덴마크 국가인권기구의 디지털권리 체크리스트를 참고할 수 있다.
<https://www.humanrights.dk/tools/digital-rights-check (접근일: 2022. 12. 1)>.

마크 국가인권기구는 2014년에도 「인권과 영향평가: 민간 부문의 맥락에서 개념적 현실적 고려(Human Rights and Impact Assessment: Conceptual and Practical Considerations In the Private Sector Context, 2014)」[18]를 펴내어 다른 영향평가(환경영향평가, 사회영향평가 등)와는 다른 인권영향평가의 고유한 성격을 정립하기 위해 노력해왔다.[19]

2020년에는 「인권영향평가 가이드 및 도구모음(Human Rights Impact Assessment: Guidance and Toolbox)」(이하 '덴마크 가이드'라 함)[20]을 개정·완간했는데 이와 함께 각 단계마다 실무자들이 참조할 수 있는 부록(Practitioner Supplements)도 제공하고 있어 매우 충실한 구성으로 평가되고 있다.[21]

이 덴마크 가이드는 기본적으로 유엔 기업과 인권 이행지침에 기반을 두고 있다. 인권영향평가에 대한 개념도 정의하고 있다. 이에 따르면, 인권영향평가는 "사업(business)의 맥락에서, 사업 프로젝트 또는 사업활동이 인권에 미치는 부정적 효과를 식별, 이해, 평가, 표명하는 과정"이다.[22] 또한 인권영향평가의 10가지 핵심기준(10 key criteria)을 제시하여,[23] 인권영향평가가 기업 경영뿐만 아니라 공공 정책에도 적용될 수 있도록 원칙과 도구를 세분화·구체화하고 있다.

이 가이드에 따른 인권영향평가(HRIA)는 사업활동을 통해 근로자, 지역사회 구성원, 소비자 등 권리주체에게 미치는 영향을 분석하는 것이며, 차별금지와 같은 인권 원칙을 영향평가의 과정에 통합한 인권기반 접근법(human rights-based approach)을 따르고 있다.

해당 인권영향평가 지침 및 도구는 기본적으로 대규모 사업 프로젝트를 모델로 하고 있지만, 맥락에 따라 소규모의 프로젝트별, 사업유형별로 적합하게 조정될 수 있으며, 인권이 환경영향평가, 사회적 영향평가, 보건영향평가 등에 통합되는 경우에도 유용하게

---

18) <https://www.humanrights.dk/publications/human-rights-impact-assessment (접근일: 2022. 12. 1)>.

19) 이준일 외(2018), 57면.

20) The Danish Institute for Human Rights(2020a).

21) 해당 부록들은 관련 현장 사례(road-testing) 및 예시 자료집(example question and resources)이다. 박준석(2022), 6면.

22) The Danish Institute for Human Rights(2020a), p.10.

23) 기존 2014년 연구에서 그 기준은 5가지로 "① 국제 인권 기준의 적용, ② 전 범위의 영향에 대한 처리, ③ 인권 기반의 절차 도입, ④ 책임(응답성)의 확보, ⑤ 영향의 심각성에 대한 평가 및 영향의 처리"였으나, 2020년 개정판에서는 "① 참여, ② 비차별, ③ 역량 강화, ④ 투명성, ⑤ 책임(응답성), ⑥ 벤치마크(국제인권 기준의 적용), ⑦ 영향의 범위(실제적·잠재적 영향 포괄), ⑧ 영향의 심각성 평가, ⑨ 영향 완화 조치, ⑩ 구제 수단에의 접근"으로 확장되었다. 박준석(2022), 7면. 확장된 10가지 핵심기준에 대하여는 The Danish Institute for Human Rights(2020a), pp.28-39을 참조할 수 있다.

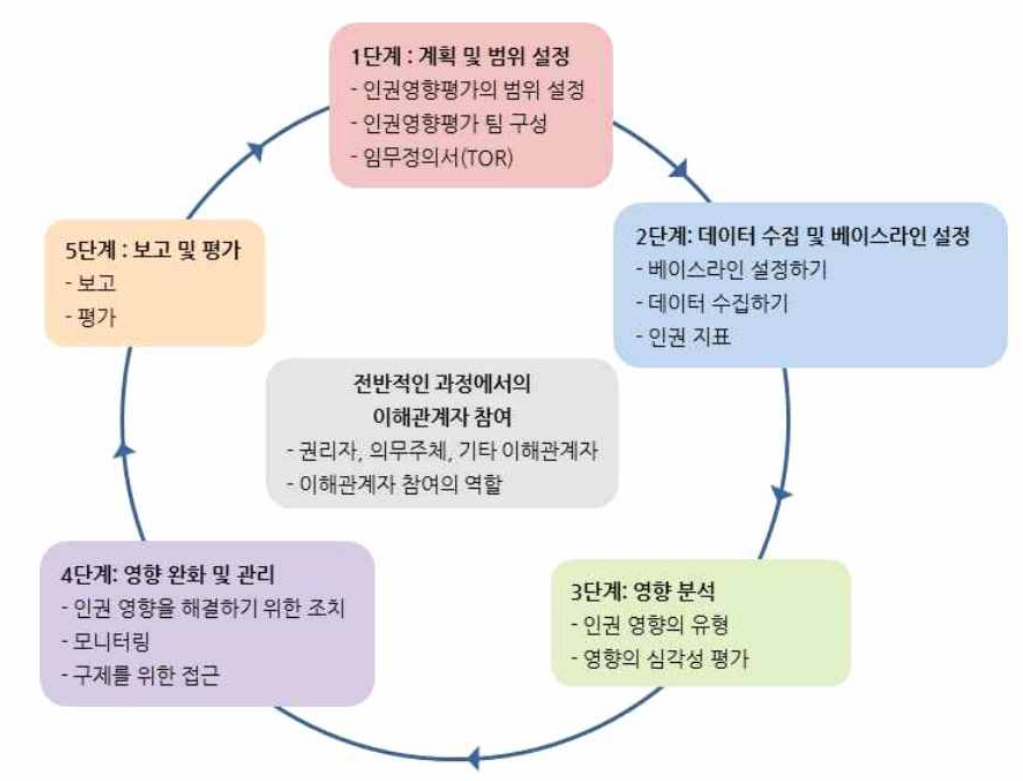
활용될 수 있다.

덴마크 가이드가 제안하는 인권영향평가의 절차는 다음과 같이 5단계로 구성된다([그림 1] 참조).

- 1단계: 계획 및 범위 설정
- 2단계: 데이터 수집 및 베이스라인 설정
- 3단계: 영향분석
- 4단계: 영향 완화 및 관리
- 5단계: 보고 및 평가

(전반적인 과정에서의 이해관계자 참여)

[그림 1] 덴마크 국가인권기구의 인권영향평가 5단계[24)]



24) 이 그림은 한국도시연구소(2019), 6면에서 인용.

이 덴마크 가이드는 다음 장에서 자세하게 살펴보게 될 2020년 덴마크 〈디지털활동 인권영향평가 지침(Guidance on Human Rights Impact Assessment of Digital Activities)〉의 절차적 기초를 제시하고 있다.

그 밖에도 참고할만한 해외 인권영향평가 가이드라인으로 영국 애버든 시 평등 및 인권영향평가 가이드,[25] 기업의 사회적 책임에 대한 가이드라인 및 절차,[26] 공공 정책에 대한 결과 기반의 인권영향평가(OPERA)[27] 등이 있다.[28]

2. 국내 인권영향평가 현황 및 사례

가. 지방자치단체

국내에서 인권영향평가 또는 인권실사는 아직 법률상 규정된 형태로 제도화되지 않았다. 인권영향평가의 제도화는 2003년 12월 국가인권위원회법 개정 논의과정에서 시도된 바 있었다.[29] 여기에서 제안된 법안의 특징은 자치법규뿐만이 아니라 "법령·정책 등의 제정·입안"을 하는 경우도 인권영향평가서를 작성하도록 하고 있다는 점이다. 그러나 이는 임기만료로 폐기되었다. 이후 2012년 4월 국가인권위원회 상임위원회는 「인권 기본조례 제·개정 권고」와 함께 「인권 기본조례 표준안」을 발표하면서 지방자치단체의 조

25) Aberdeen City Council(2008). Equality and Human Rights Impact Assessment : the Guide.

26) BSR(Business for Social Responsibility)(2013). Conducting an Effective Human Rights Impact Assessment: Guidelines, Steps, and Examples.

27) Center for Economic and Social Rights(2012). Assessing Fiscal Policies from a Human Rights Perspective.

28) 기타 △건강권 인권영향평가와 관련한 Gostin & Mann 모델 및 HeRWAI 모델, △유럽연합의 「무역 관련 정책 이니셔티브를 위한 영향평가 내 인권영향분석에 관한 지침」(Guidelines on the analysis of human rights impacts in impact assessments for trade-related policy initiatives), △2011년 「무역 및 투자협정의 인권영향평가에 관한 지침원칙」(Guiding principles on human rights impact assessments of trade and investment agreements), △유엔-유니세프의 '아동권리 툴킷: 개발협력과 아동권리 통합'('Child Rights Toolkit: Integrating child rights in development cooperation'), △UNICEF와 덴마크인권연구소가 공동으로 개발한 '영향평가에서의 아동권리'(Children's Rights In Impact Assessments), △북유럽신탁기금(Nordic Trust Fund)과 세계은행(World Bank)이 공동으로 연구하고 2013년 발표한 「인권영향평가에 관한 연구: 문헌 및 개발에 관한 다른 유형의 평가와의 차이점 및 관련성 검토」(Study on Human Rights Impact Assessments A Review of the Literature, Differences with other Forms of Assessments and Relevance for Development) 등이 있다. 이에 대한 간략한 소개로는 이준일 외(2018), 62면 이하; 한국도시연구소(2019), 7면 이하 참조.

29) 이 과정과 관련 개정안의 내용에 대하여는 박준석(2022), 8면 참조.

례 제·개정이나 정책 수립에서 인권영향평가 제도를 도입할 근거를 마련하였다. 이후 여러 광역 및 기초지방자치단체에서 인권 관련 조례를 제정하고 인권영향평가를 실시하게 되었다. 아래 표에서 볼 수 있는 것처럼 모든 광역 지자체가 인권 관련 조례를 제정하였고, 이 중 70% 넘게 이 조례에 인권영향평가 제도를 규정하였다.[30]

[표 1] 광역지방자치단체 인권영향평가 제도 도입 현황[31]

광역자치단체	인권에 관한 조례	인권영향평가	인권영향평가 도입
서울특별시	인권기본조례	제8조	2016년
부산광역시	인권기본조례	제14조의6	2019년
대구광역시	인권보장 및 증진에 관한 조례	×	-
인천광역시	시민인권보장 및 증진에 관한 조례	제9조 제2항	2019년
광주광역시	인권보장 및 증진에 관한 조례	제30조	2012년
대전광역시	인권보호 및 증진 조례	×	-
울산광역시	인권기본조례	제8조	2020년
세종특별자치시	인권보장 및 증진에 관한 조례	×	-
경기도	인권보장 및 증진에 관한 조례	제9조	2021년
강원도	인권보장 및 증진에 관한 조례	×	-
충청북도	인권보장 및 증진에 관한 조례	제9조	2018년
충청남도	인권기본조례	제9조	2018년
전라북도	도민 인권보호 및 증진에 관한 조례	제8조의2	2020년
전라남도	인권기본조례	제15조	2015년
경상북도	인권보장 및 증진에 관한 조례	×	-
경상남도	인권보장 조례	제7조의2	2021년
제주특별자치도	인권보장 및 증진에 관한 조례	제20조	2020년

30) 박준석(2022), 12면.
31) 박준석(2022), 12-13면에서 인용.

이 중 광주광역시는 광역지방자치단체 최초로 2017년 인권영향평가 제도를 도입하였는데, 자치법규(조례, 규칙) 제·개정, 정책·사업 수립·시행 과정에서 시민의 인권침해 및 인권증진 가능성 정도를 사전에 분석·평가하여 행정이 인권증진에 기여할 수 있도록 인권영향평가를 도입하였다고 밝히고 있다.[32]

광주광역시의 인권영향평가는 인권부서에서 주관하며, 사업부서가 제출한 관련 자료를 기초로 외부 전문가 또는 전문기관에 의뢰하여 검토·평가를 진행하고, 평가결과를 사업부서에 통보하여 개선·권고사항에 대한 이행도를 지속 점검하는 절차로 진행된다.[33]

기초지방자치단체들의 경우, 인권조례를 제정하고 인권영향평가를 실시하는 지방자치단체의 비율은 그다지 높지 않으며, 자치법규로 규정하였더라도 실제로 인권영향평가를 실시하는 지방자치단체는 극히 소수에 불과하다.[34]

기초지방자치단체 중 실제로 인권영향평가를 실시하고 있는 모범사례라 할 수 있는 수원시의 경우, ‘수원시 자치법규 인권영향평가 매뉴얼(2018)’, ‘수원시 공공건축물 인권영향평가 실행방안 연구보고서(2019)’ 등을 발표하고, 자치법규, 정책(계획), 공공건축물에 대한 인권영향평가를 실시하고 있다.[35]

수원시의 인권영향평가는 다음과 같이 평가시기, 평가주체, 평가방법, 평가내용을 제시하고 있다([그림 2] 참조).[36] 수원시는 인권영향평가의 절차 또한 자치법규, 정책, 공공건축물 분야로 나누어 시행하고 있다([그림 3] 참조).

---

32) 광주광역시(2022), 250면.

33) 추진상황을 살펴보면, 조례, 규칙 제·개정(2018년: 평가 17건, 권고 9건/2019년: 평가 97건, 권고 11건/2020년: 평가 135건, 권고 8건/ 2021년: 평가 175건, 권고 10건), 공공건축물(빛고을안전체험관, 하남지구 시립도서관, 장애인 수련시설), 투표소 인권모니터링(제7회 6·13지방선거), 사회복지시설 규정-편람, 공공기관 주관행사에 대한 인권영향평가를 실시하였다.

34) 수원시와 서울시 성북구 등의 경우에만 실제로 인권영향평가를 실시하고 그밖의 대부분 기초지방자치단체는 투표소 인권영향평가 등 일부 정책을 대상으로 간헐적으로 시행하고 있는 실정이다. 이충은·노진석(2018), 223면 참조.

35) <https://www.suwon.go.kr/sw-www/www02/www02-11/www02-11-11/www02-11-11-02.jsp (접근일: 2022. 12. 1)>.

36) 수원시의 인권영향평가의 자세한 내용은 이충은·노진석(2018), 229면 이하 참조.

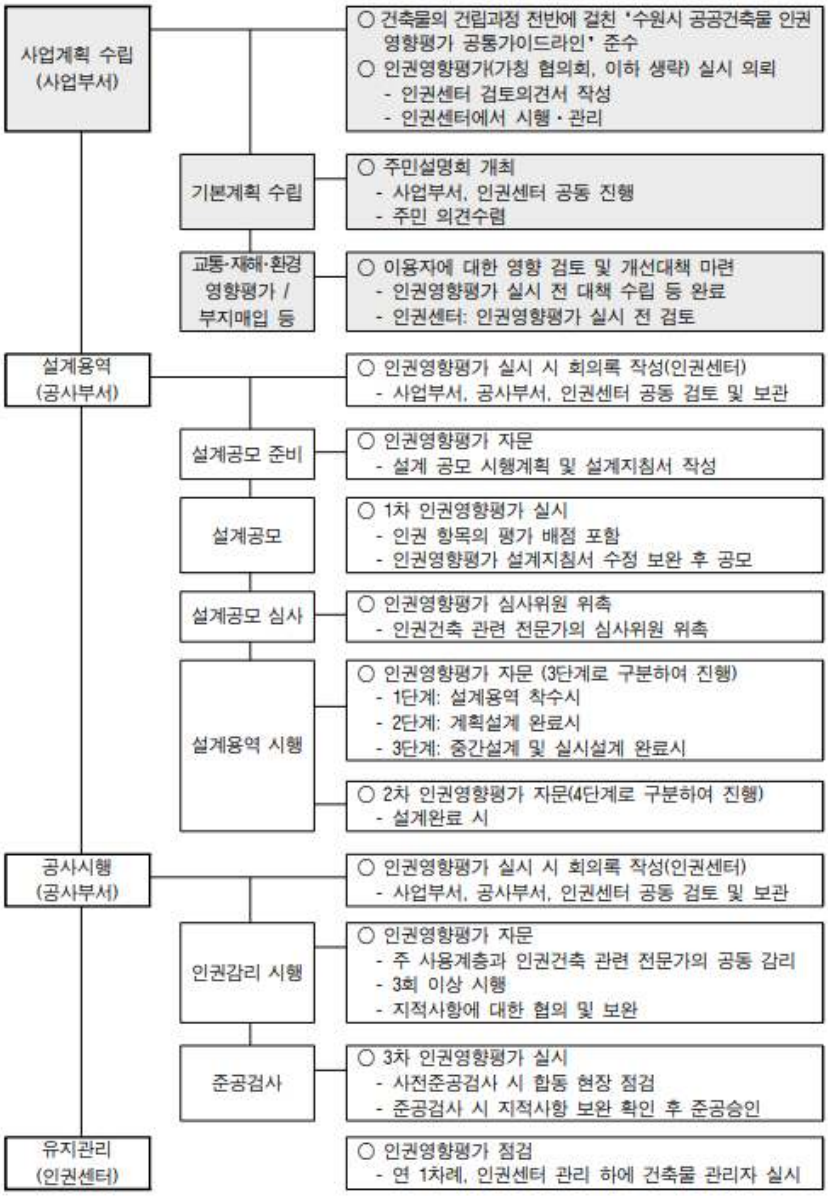
[그림 2] 수원시 인권영향평가의 개요

	자치법규(조례·규칙)	정책(계획)	공공건축물
평가시기	조례·규칙심의회 심의·의결 전	해당 정책 확정 전	해당 사업계획 수립 시
평가주체	✓ 인권센터 ✓ 인권정책교육소위원회	✓ 인권센터 ✓ 인권정책교육소위원회	✓ 인권센터 ✓ 인권영향평가협의회
평가방법	1차 : 자체 점검표(담당부서) 2차 : 인권센터 등 평가	1차 : 자체 점검표 (담당부서) 2차 : 인권센터 등 평가 ※ 일반 / 특정평가 구분	외부 전문평가단 자문
평가내용	✓ 4대 평가영역 : 인권침해, 침해구제, 참여권, 인권증진 ✓ 6대 평가기준		✓ 주민의견 수렴 ✓ 인권을 반영한 설계 지침 마련

[그림 3] 수원시 인권영향평가의 절차

이 중에서 공공건축물 인권영향평가와 관련하여 ‘수원시 공공건축물 인권영향평가 실행방안 연구’ 보고서에서는 그 시행절차를 다음의 [그림 4]와 같이 제시하고 있다.[37]

[그림 4] 수원시 공공건축물 인권영향평가 시행절차



37) 안국진(2018), 27면.

나. 공공기관 및 사기업

제3차 국가인권정책기본계획(2018-2022)은 ‘인권경영의 제도화’를 기본계획의 과제로 제시하면서 “대한민국 영토 또는 관할권 내의 모든 기업은, 그 규모나 활동 장소에 상관없이, 유엔 기업과인권 이행지침에서 정의된 기업의 인권 존중 책임을 이행”한다고 언급하고 있다. 그러나 앞서 살펴본 바와 같이 인권영향평가 또는 인권실사는 법률상 제도화되어 있지 않다. 따라서 기업은 공·사를 불문하고 인권영향평가의 의무적 실시가 강제되지 않는다.

공공기관의 경우 인권영향평가(인권실사)는 경영평가제도의 일환으로 간접적 강제가 이루어지고 있다.[38] 국가인권위원회는 “인권경영 가이드라인 및 체크리스트”(2014년 9월), “공공기관 경영평가 제도개선”(2016년 2월) “공공기관 인권경영 매뉴얼 도입”(2018년 3월) 등을 권고한 바 있고, 2022년 7월에는 “공공기관의 인권경영 강화를 위한 ‘인권경영 보고 및 평가지침’ 적용”을 권고하였다. 여기에서는 “30개 정부 부처 및 17개 광역자치단체장에게, 경영평가 대상인 산하 공공기관이 인권위가 마련한 ‘인권경영 보고지침’에 따라 인권경영의 결과를 보고 및 공시하도록 하고, 향후 기관 경영평가 시 위 지침에 따라 독립적인 항목으로 인권경영을 평가할 것”을 권고하고 있다.

정부부처와 광역지방자치단체가 위 권고를 수용하면서, 결과적으로 여러 공공기관이 인권경영 매뉴얼에 따라 인권정책선언을 수립·공개하고, 인권경영위원회 설립, 인권영향평가 실시 등 인권경영 체계를 구축하여 경영평가를 받고 있다.

「공공기관 인권경영 매뉴얼」(2018)은 유엔의 기업과 인권 이행지침에 따라 인권영향평가를 “인권리스크를 평가하기 위해 기관(기업)이 사업 관계의 결과로 또는 기업의 활동으로 인해 인권에 미칠 수 있는 실제적·잠재적인 인권리스크를 파악하고 평가하는 절차”로 정의한다. 또한 인권영향평가를 “기관(기업)운영 인권영향평가”와 “주요사업 인권영향평가”로 구분한다. 기관(기업)운영 인권영향평가는 “기업 활동 전반을 대상으로 실시하는 평가로, 인권경영 체제, 고용, 노동권, 산업안전, 공급망, 현지주민 등 포괄적인 분야를 대상으로 평가를 실시”하는 것을 말하며, 주요사업 인권영향평가는 “기관

38) 김동현(2022), 123면 참조.

(기업)이 추진하는 특정 사업을 대상으로 실시하는 평가로, 해당 사업이 인권에 미치는 부정적 영향을 사전에 파악하고 분석하여 이를 예방하거나 최소화하기 위한 평가"를 말한다.[39]

기관(기업)운영 인권영향평가의 시행절차는 다음과 같다. ① 기관(기업)운영 인권영향평가 실시 계획 수립, ② 인권경영 가이드라인 및 체크리스트 교육, ③ 인권경영위원회 평가 자료 제출, ④ 인권경영위원회 평가 및 결과 보고서 작성, ⑤ 최고경영진 보고 및 공개.

주요사업 인권영향평가도 이와 유사한 절차로 진행된다. ① 주요사업 인권영향평가 실시 계획 수립, ② 주요사업 인권영향평가 지표 마련, ③ 주요사업 인권영향평가 지표 교육, ④ 인권경영위원회 평가 자료 제출, ⑤ 인권경영위원회 평가 및 결과 보고서 작성, ⑥ 최고경영진 보고 및 공개.

현재 경제·인문사회연구회, 한국환경공단, 한국소비자원, 통일연구원, 한국노동연구원, 한국환경연구원, 한국중부발전, 한국마사회, 전력거래소, 한국농수산식품유통공사, 경기주택도시공사, 한국교통연구원, 한전케이페에스주식회사, 국토연구원, 광명도시공사 등 여러 공공기관에서 인권경영의 일환으로 인권영향평가를 실시하고 결과를 보고·공시하고 있다.

예를 들어, 2019년부터 인권영향평가를 실시하여 결과를 발표하고 있는 한국환경공단의 경우, 기관운영 인권영향평가와 주요사업 인권영향평가(2021년의 경우, 5개 주요사업, 1개 지역본부 선정)를 실시하고 있으며, 담당부서 자체평가(1차, 지표별 담당자가 체크리스트 점검 및 증빙자료 제출)와 전문가 검증(2차, 증빙자료 적절성 검토 및 인터뷰 실시)의 방법으로 평가를 진행한다. 다음에서 보는 바와 같이 국가인권위원회의 「공공기관 인권경영 매뉴얼」을 기반으로 평가절차를 구성하여 시행하고 있다.

39) 국가인권위원회(2018), 14면.

[그림 5] 한국환경공단 인권영향평가 절차[40)]

구분	내용
[1단계] 계획 수립	- 인권리스크 진단이 필요한 2021년 주요사업 인권영향평가 대상사업 선정(주요사업 5개+지역본부 1개) - 인권영향평가 대상, 일정, 방법 등 실시 계획 수립
[2단계] 지표 설계	- 기관운영 평가지표 고도화 - 주요사업 평가지표 개발
[3단계] 평가 교육	- 지표별 담당부서 대상 인권영향평가 관련 실무교육 실시
[4단계] 평가 실시	- 지표별 담당부서 자체평가(체크리스트 점검 및 증빙자료 제출) - 외부 이해관계자 의견 수렴(설문조사 통한 주요 협력사 의견 수렴) - 전문가 검증(증빙자료 적절성 검토 및 인터뷰 실시)
[5단계] 결과 분석	- 평가 결과 도출 및 결과보고서 작성 - 평가결과 최고경영진 보고 및 홈페이지 등 공개

* 국가인권위원회「공공기관 인권경영 매뉴얼(2018.8.)」기반으로 실시

한편, 법무부는 기업과 인권 이행지침의 핵심내용인 인권실사의 과정이나 구제절차의 이해를 돕기 위해「기업과 인권 길라잡이」(2021)를 발간하면서, 인권영향평가를 포함하는 인권실사의 포괄적 절차를 제시하였다. 이에 따르면, 인권실사는 "기업이 끼치는 부정적 인권영향을 식별하고 방지·완화하며, 이러한 노력을 대내외에 천명하는 포괄적 절차"를 말한다. 여기에서 말하는 포괄적 절차는 "인권실사의 대상 및 우선순위 선정, 실재적·잠재적 인권리스크 식별과 평가(인권영향평가), 그 결과를 기업운영과 활동 전반에 반영하고 실천하는 과정, 취해진 조치의 효과성에 대한 모니터링, 그리고 이 모든 절차에 대한 정보를 공개하는 단계로 구성"되며, "모든 단계에서 이해관계자와의 소통과 협력"이 중요하다. 또한 인권실사의 절차와 방법은 일률적인 것이 아니라, 기업의 규모 및 가용자원, 기업이 대내외적으로 직면한 인권리스크의 수준 및 심각성 등에 따라 유연하게 적용할 수 있다고 한다.[41)]

40) 한국환경공단(2022), 1면.
41) 법무부(2021), 39면 참조.

제2절 시사점

이상에서 살펴본 바와 같이, 국내외의 인권영향평가 제도는 비록 초보적이지만 규범적 선언의 수준에서 벗어나 구체화·제도화를 논의하는 단계로 나아가고 있는 것으로 보인다. 일부 국가에서는 인권실사와 그 핵심도구인 인권영향평가를 제도화하여 의무적으로 실시하기 시작하였고, 그 절차와 기준도 상대적으로 구체적이다. 이러한 동향에 비추어 우리나라에서 인공지능 인권영향평가의 개발 및 시행에 있어 고려해 볼 수 있는 시사점은 다음과 같다.

국내외 동향을 통해 공통적으로 확인할 수 있었던 것은 인권영향평가의 절차 또는 평가단계가 유사하게 수렴되고 있다는 점이다. 인권영향평가의 절차는 대체로 사전 준비 단계(계획 및 정보수집), 평가 및 분석 단계, 영향 완화 및 관리 단계, 보고 및 점검 단계 등으로 구성된다. 이 모든 과정에서 이해관계자의 참여가 강조되기도 한다.

다음으로 인권실사 및 인권영향평가 제도가 자율적 실시에서 의무적 실시로 제도화·법제화되고 있다는 점이다. 미국, 영국, 호주, 프랑스, 독일, 네덜란드, 노르웨이 등은 인권실사를 법체계 내에 제도화하였으며, 유럽연합의 일부 지침과 규칙도 인권실사를 제도화한 것으로 평가되고 있다.[42] 하지만 우리나라 인권영향평가는 지방자치단체의 조례에 근거하여 제도화되기 시작하여 경영평가의 일부로 간접강제되는 등 제도적으로 불완전한 형태를 취하고 있다.

인권영향평가의 제도화와 관련하여 취할 수 있는 입법전략은 ① 현행 유지, ② 구속력 없는 지침 신설, ③ 기존의 기업의 보고의무에 초점을 맞춘 규제를 더욱 강화하는 방식의 새로운 규제 신설, ④ 법적 주의 의무 기준으로 실사를 의무화하는 방식의 새로운 규제의 신설 방안 등이다.[43] 이와 관련하여 ④를 채택하는 경우 ②와 ③을 채택하는 경우보다 추가 비용이 증가할 것이지만, 사회적으로 더 긍정적 영향을 미칠 것으로 예상된다고 본다.

이와 관련하여, 인권실사(인권영향평가)의 본질이 이해관계자(특히 인권피해자)의 인권

---

42) 김동현(2022), 121면.

43) 이는 인권실사의 의무화와 관련하여 국가가 택할 수 있는 규제 전략으로, 유럽의회가 의뢰하여 실시한 연구의 결과에 따른 것이다. 이에 관하여는 김동현(2022), 118면 이하 참조.

침해 방지를 위한 사전예방적 조치라는 점에서 기업의 재무적 손상여부를 강조하는 인권 위험관리와 구분되어야 하며, “인권실사의 제도화”의 의미를 ‘인권정책선언, 이해관계자의 참여가 보장된 인권영향평가와 공개’를 포함하는 것으로 전제할 때, 공기업이나 공적 자금을 이용하는 사기업, 그리고 상장기업 정도의 대기업에 한정하는 사기업의 경우 인권실사를 법률적으로 요구할 수 있다는 주장도 경청할만하다.[44]

특히 인공지능에 대한 인권영향평가에서 명확한 법률적 근거를 마련할 필요성이 있다. 인공지능에 대한 인권영향평가는 개별 인공지능 기술의 개발 및 도입에 대한 영향평가와 인공지능 관련 입법 그리고 정책에 대한 영향평가로 구분해 볼 수 있다. 이상적으로는 체계적 입법과 정책 설계를 위해 후자의 인공지능 영향평가가 먼저 제도화되는 것이 바람직하다. 그러나 인공지능 입법 및 정책에 관한 인권영향평가를 제도화하기에는 아직 관련 연구가 성숙하지 않았고, 더욱이 신생 기술로서 인공지능이 인권에 미치는 영향이 불확정적이고 관련 인권 관련 쟁점도 명확하지 않은 실정이다. 더욱이 우리나라는 법령에 대한 인권영향평가도 입법되지 않았음은 앞서 살펴본 바와 같다.

따라서 인공지능 기술의 개발 및 도입에 대한 인권영향평가의 제도화를 먼저 시도할 필요가 있다. 문제는 인공지능 기술에 대한 인권영향평가가 평가대상에게 법적 의무를 부과하고 일정한 비용부담을 발생시킨다는 점이다. 따라서 초보적인 수준에서라도 인공지능 기술에 대한 인권영향평가의 제도화를 위해서는 법률적 근거 마련이 필요하다.

이와 관련하여 이하에서 자세하게 설명하겠지만 인공지능 인권영향평가의 시행을 위해 기존에 법제화되어 있는 영향평가제도를 활용하는 방안은 일정한 한계를 갖는다. 현재 우리 법제에는 수많은 영향평가(분석)가 난립해 있는 상황이다. 행정기본법상 이제 막 도입된 입법영향분석을 별론으로 하고 이른바 통합영향평가(integrated impact assessment) 체계가 구축되지 않은 우리나라 실정에서 수많은 영향평가 제도들 중 인권영향평가의 성격과 위상 또한 아직은 불분명하다. 이에 더해 인공지능 인권영향평가가 이러한 영향평가제도와 정합적 관계를 맺기 위해서는 여러 단계의 입법과정이 필요함을 알 수 있는 대목이다.

또한 인공지능 인권영향평가와 친화성이 예상되는 규제영향평가(분석), 기술영향평가(분석), 사회적 영향평가, 입법영향평가(분석), 성별영향평가, 아동정책영향평가, 환경영향

44) 이상수(2015), 95면 이하 참조.

평가 등은 평가대상이 제한적일 뿐 아니라 인공지능 영향평가의 제도화에 있어 부족한 제도들이다.

예를 들어 과학기술기본법상의 기술영향평가는 요건상으로 인공지능 인권영향평가도 가능하지만, 인공지능 기술에 잠재하는 인권위험을 체계적이고 집중적으로 평가하고 분석하는 틀로는 부족하며, 평가도 간헐적이고 일회적으로 수행되고 있음을 고려하여야 한다. 지능정보화 기본법상 사회적 영향평가의 경우, 인공지능을 대상으로 하는 영향평가라는 점에서 의의를 찾을 수 있지만, 평가기준으로 인권이 명시적으로 규정되어 있지 않으며 평가대상도 포괄적이고 소관부처도 평가대상마다 달라진다는 문제점이 있다. 개인정보 보호법상 개인정보 영향평가의 경우, 그 목적상 개인정보 침해 사안을 주로 하는 영향평가이며, 공공기관을 주된 대상으로 한다는 점에서 인공지능 인권영향평가의 일부 기능만을 담당할 수밖에 없다는 한계를 갖는다.

따라서 기존 국가인권위원회법 개정이나 새로운 입법을 통해 인공지능 인권영향평가를 도입하는 방안을 단계적으로 검토할 필요가 있다.

마지막으로 우리나라의 인권영향평가, 특히 인공지능 인권영향평가 도입에서 강조되어야 할 지점은 "평가의 방법론"이다.[45] 인권영향평가의 국내 도입과정에서 초창기 인권영향평가의 대상이 자치법규였다는 점은 일종의 방법론적 편향을 가져왔다. 다시 말해 주로 지방자치단체의 자치법규를 대상으로 하면서. 인권영향평가가 조례나 규칙에 대한 사법심사와 유비되고, 인권영향평가의 방법론도 완화된 심사척도를 모사하는 형태를 취하게 되었다. 이에 따라 인권에 미치는 부정적 또는 긍정적 영향을 데이터에 기반하여 평가하는 '과학적 방법론'은 크게 주목받거나 발달하지 못했다. 앞서 언급한 입법 그리고 정책에 대한 인권영향평가 중 입법을 대상으로 하는 영향평가는 물론 규범체계와 규범정합성을 축으로 하는 규범적 방법론이 평가의 주를 이루어야 할 것이다. 그러나 정책에 대한 인권영향평가나 기술 등에 대한 인권영향평가는 증거와 데이터에 기반한 과학적 방법이 합리성과 객관성을 담보할 것이다.

특히 인공지능 인권영향평가는 인공지능이라는 미지의 기술을 대상으로 관련 입법, 정책, 사업, 기술이 인권에 미치는 영향을 측정하고 평가하는 작업이다. 따라서 다른 영향평가보다 과학적 방법의 주를 이루는 평가 분야라 할 수 있다. 물론 인권의 가치를 놓치

45) 이에 대해서는 박준석(2022), 13면 이하 참조.

지 않는 규범적 토대가 전제되어야 하겠지만(value added), 인공지능 인권영향평가에서만큼은 과학적 방법이 무엇보다 강조되어야 한다.

인공지능 인권영향평가 도입과정에서 숨어 있는 쟁점은 인권감수성과 함께 디지털 문해력(literacy)의 자질을 갖춘 전문인력의 존재 여부이다. 공·사, 기업 내·외부에서 이러한 전문인력을 확보하지 못한다면 인공지능 인권영향평가의 성공을 장담할 수 없을 것이다.

인공지능 인권영향평가의 절차적 핵심이 정보수집(data collection)과 분석(analysis)이라는 점을 감안한다면, 인공지능이 야기할 수 있는 다양한 인권 문제에 관해 증거에 기초하여 합리적으로 평가할 수 있는 자질을 갖춘 전문인력과 전문기관의 필요성은 더욱 높아질 것이다. 인공지능 기술과 그로 인한 인권영향을 측정할 수 있는 역량을 강화하기 위해서는 인권영향평가의 과학적 방법론, 데이터와 증거에 기반한 평가방법론을 세심하게 고민할 필요가 있다.

제3장 인공지능 영향평가 사례

인공지능이 사람의 인권과 안전에 미치는 영향에 대한 우려가 커짐에 따라, 그 위험을 예방하고 완화하기 위한 방안으로 영향평가 제도가 주목받고 있다. 최근 몇 년 간 여러 연구자들과 시민사회에서 다양한 유형의 영향평가도구를 제안하여 왔다. 예를 들어 뉴욕대학교 AI Now 연구소는 2018년 <알고리즘영향평가: 공공기관 책임을 위한 실용적인 프레임워크(Algorithmic Impact Assessments: A practical framework for public agency accountability)>[46] 제하의 영향평가 모델을 개발하였다. 같은 해 존스 홉킨스 대학교 정부우수성센터(GovEx), 샌프란시스코 시군, 하버드 케네디 스쿨 데이터스마트 프로젝트 및 비영리단체 데이터커뮤니티 DC는 공동으로 <윤리 및 알고리즘 툴킷(Ethics & Algorithms Toolkit)>[47]을 개발하고 공공부문에서 의사결정 지원 알고리즘을 사용할 때 이 툴킷을 사용할 것을 권고하였다. 2020년에는 시민단체 미국시민자유연합(ACLU)이 질문지 형식의 <알고리즘 형평성 툴킷(Algorithmic Equity Toolkit, AEKit)>[48]을 개발하였다.

이러한 논의에 힘입어 최근 유럽연합, 캐나다, 영국, 미국 등 주요 국가들은 인공지능 및 알고리즘에 대한 영향평가 제도를 도입해 왔다. 특히 세계 각국은 공공부문 인공지능과 민간부문 고위험 인공지능이 사람들에게 미치는 부정적 영향에 주목하고 이를 예방적으로 해결하기 위한 방안으로 다양한 인공지능 및 알고리즘 평가를 제안하고 있다.

인공지능 기술의 도입 초창기, 공공부문은 산업계에 비하여 전문성 및 리더십이 부족하였고 일명 인공지능 '블랙박스' 문제[49]로 인하여 인공지능 기술에 대한 정책적 개입에 소극적이었던 것이 사실이다. 그러나 사람을 대상으로 고용, 사회복지 급여, 재판, 입시 등 중요한 의사결정을 내리는 데 사용되는 공공부문 인공지능이 민간부문의 비공개

---

46) AI NOW(2018). Algorithmic Impact Assessments: A practical framework for public agency accountability. <https://ainowinstitute.org/aiareport2018.pdf(접근일: 2022. 8. 15)>.

47) GovEx, the City and County of San Francisco, Harvard DataSmart, Data Community DC(2018). Ethics & Algorithms Toolkit. <https://ethicstoolkit.ai/(접근일: 2022. 8. 15)>.

48) American Civil Liberties Union(2020). Algorithmic Equity Toolkit. <https://www.aclu-wa.org/AEKit(접근일: 2022. 8. 15)>.

49) 머신러닝으로 방식으로 작동하는 인공지능 시스템은 설명하기 어렵거나 불가능한 방식으로 패턴을 식별하고, 설명하기 어렵거나 불가능한 처방을 내릴 수 있다. 이를 흔히 '블랙박스' 문제라고 한다. UNITED NATIONS(2021), para.20. 참조.

알고리즘에 의존하면서 그 의사결정에 대하여 설명하지 못하거나, 인종, 지역 등에 편향적이거나 차별적인 의사결정을 내린 데 따른 논란이 계속 불거져 왔다.[50] 이에 인공지능의 투명성, 책임성 및 책무성 등을 확보해야 할 필요성이 지적되었고, 특히 인공지능 개발과 활용 과정에서 소비자 보호나 개인정보 보호 등에 관한 법률 위반이 발생하는 데 대하여 규제기관의 대응이 요구되어 온 상황이다.

관련하여 영국 경쟁시장 감독기구(CMA), 방송통신 규제기구(Ofcom), 개인정보보호 감독기구(ICO), 금융감독기구(FCA)은 2020년 7월 공동으로 디지털규제협력포럼을 구성하고 2022년 4월 28일 알고리즘 감사 현황 및 규제기관 역할에 대한 보고서를 발간하였다.[51]

[표 2] 영국 디지털규제협력포럼의 알고리즘 감사 분류

	거버넌스[관리] 감사 (Governance audit)	경험적 감사 (Empirical audit)	기술 감사 (Technical audit)
설명	올바른 거버넌스 정책을 준수하고 있는지 평가	입력 또는 출력을 사용하여 알고리즘의 효과를 측정	알고리즘의 '보닛 아래'에서 데이터, 소스 코드, 방법론을 살펴봄
방법	영향평가, 준수 감사(투명성 감사 포함), 적합성 평가	스크래핑 감사, 미스터리 쇼핑객 감사	코드 감사, 성능 테스트, 공식 검증
사례	EU 인공지능법(안)은 인공지능 고위험 애플리케이션에 대한 적합성 평가를 의무화함	프로퍼블리카는 COMPAS에 의한 재범[위험성 평가] 알고리즘에 대한 조사를 수행할 때 예측된 재범률을 2년 동안 실현된 재범률과 비교함	내부 코드 피어 리뷰[동료 평가]가 구글의 업무흐름 개발에서 일반적인 관행이 됨

50) 미국 교육청의 교사평가 알고리즘, 폴란드와 네덜란드 사회복지 인공지능 등에 대하여 적법절차 또는 투명성 의무를 위반하였다는 각국 법원의 판결이 있었다. 미국에서 재판절차에 쓰이는 COMPAS는 인종 편향 논란이 불거지기도 하였다. 김기중 외(2021), 제7장 참고. 한편, 유엔 인권최고대표는 성적 예측 인공지능이 "공립학교와 가난한 지역의 학생들을 차별하는 결과를 낳았다"고 지적했다. United Nations(2021) para.2 참고. 이 인공지능은 2020년 영국 정부가 도입하였다가 거센 비판을 받고 철회한 바 있다. 한겨레(2020. 8. 24). "AI가 준 학점, 가난한 학생을 차별했다" 참고.

51) Digital Regulation Cooperation Forum(2020). Auditing algorithms: the existing landscape, role of regulators and future outlook.

이 보고서는 알고리즘 감사(auditing)를 “알고리즘 처리 시스템을 검토하기 위한 다양한 접근방식”으로 폭넓게 소개하면서 그 방법을 [표 2]와 같이 분류하였다. 여기서 영향평가(Impact assessment)의 경우 관리적 측면에서 거버넌스 감사 방법에 속하며, 다른 알고리즘 감사 방법에 비하여 시장에 출시하거나 서비스를 제공하기 전에 수행하고 그 주요사항을 공개할 것이 요구된다는 공통점이 있다.

인공지능 영향평가 제도에 대한 제안은 크게 위험기반 접근(risk-based approach)에서 제안하는 영향평가와 인권기반 접근(rights-based approach)에서 제안하는 영향평가가 각각 논의되어 왔다. 위험기반 접근에서 제안하는 영향평가는 위험 수준이 높아질수록 요구사항을 엄격하게 적용하고 주로 고위험 규제에 초점을 둔다. 캐나다의 경우 2019년 공공조달을 담당하는 재정위원회 훈령으로 공공기관 인공지능에 대한 영향평가를 법규화하여 실시 중이고, 영국은 공공조달 지침과 국민 보건 서비스(NHS)에서 영향평가를 실시하고 있다. 유럽연합은 2021년 4월 21일 유럽연합 집행위원회가 발의한 인공지능법(안)에서 인공지능의 위험 수준을 △금지되는 위험, △고위험, △제한적인 위험, △최소 위험으로 구분하면서 특히 고위험 인공지능의 제공자에게 위험평가와 사전적합성평가를 의무화하였다. 우리나라에서도 과학기술정보통신부가 2021년 5월 14일 발표한 〈신뢰할 수 있는 인공지능 실현전략〉에서 「지능정보화 기본법」에 기반하여 인공지능 영향평가와 고위험 인공지능 기준을 도입하겠다는 방침을 밝힌 바 있다.

한편, 국제인권기구는 신기술 문제에 대한 인권기반 접근을 강조해 왔다. 유엔 인권최고대표는 2021년 9월 발표한 〈디지털시대 프라이버시권〉 보고서에서 “인권 기반 접근법은 사회가 기술적 발전의 혜택을 극대화하는 동시에 위해를 방지하고 제한할 수 있는 방법을 파악하는 데 도움이 되는 도구를 제공한다.”고 지적하였다. 또한 인공지능에 대한 인권 기반 접근법의 요구사항에 대해서도 다음과 같이 설명하였다. 우선 인권 기반 접근법은 평등과 차별금지, 참여와 책무성, 지속 가능한 개발 목표와 유엔 기업과 인권 이행지침의 중요 원칙을 포함한 여러 가지 핵심 원칙의 적용을 요구한다. 또한, 합법성, 정당성, 필요성 및 비례성의 요건이 인공지능 기술에도 일관되게 적용되어야 한다. 더불어, 인공지능은 가용성, 경제성, 접근성 및 품질이라는 핵심 요소를 달성함으로써 경제적, 사회적 및 문화적 권리의 실현을 촉진하는 방식으로 배치되어야 한다. 마지막으로 인공지능 사용과 관련된 인권 침해와 남용 피해를 입은 사람들은 효과적인 사법적 및

비사법적 구제수단을 이용할 수 있어야 한다.[52]

특히 인권기반 접근법은 인공지능이 인권에 미치는 위험을 예방하고 완화하기 위하여 인권영향평가 시행을 비롯한 인권실사를 요구해 왔다. 인권영향평가는 개인정보 등 특정한 분야에 미치는 영향을 넘어 차별받지 않을 권리, 공정한 재판을 받을 권리, 집회시위의 자유 등 모든 인권에 미치는 영향을 포괄적으로 평가하고 대응할 수 있다는 점에서 인공지능에 대한 적용이 모색되고 있다.

인권영향평가는 그 잠재적 침해의 심각도(severity)를 고려하고 심각성이 높을수록 더욱 신속하고 엄격한 조치를 취하도록 한다. 그런 점에서 인권기반 접근 또한 위험기반 접근을 반영하고 있다고도 볼 수 있다. 유럽평의회 인공지능 특별위원회는 2021년 발표한 〈인공지능 시스템에 대한 인권・민주주의・법치 영향평가〉 초안에서 인권영향평가는 인권에 미치는 위험을 평가한다는 점에서 위험기반 접근을 적용하고 있다고 보았다.[53]

그러나 덴마크 국가인권기구는 개인정보보호 영향평가(Data Protection Impact Assessment, DPIA) 등 다른 위험기반 접근법과 인권영향평가가 상호반영 또는 결합적인 방식으로 실시될 수는 있으나, 그 평가 기준(benchmarks) 측면에서 완전히 대체될 수 없다고 지적하였다. 인권영향평가는 특정 인권 항목이 아니라 모든 인권 항목에 대하여 전체적, 포용적, 포괄적 접근방식을 강조하고 영향을 받는 당사자의 참여와 이들에 대한 투명성을 중시하는 국제인권규범을 기준으로 삼는다는 점에서 다른 접근법과 다르다는 것이다.[54] 무엇보다 인권영향평가는 기업의 내부적인 위험이 아니라 사람들, 환경, 사회 등 '기업 외부차원'에서 발생하는 부정적 영향으로서 위험을 식별하고 대처하고자 한다는 점에서 여타의 위험평가와 차이가 있다.[55]

제1절은 위험기반 접근법으로 알고리즘영향평가를 제안하였거나 시행하고 있는 유럽연합, 캐나다, 영국, 미국 등의 사례를 살펴보고, 제2절은 인권기반 접근법으로 인권영향평가를 제안하였거나 시행하고 있는 유엔 인권규범, 유럽평의회, 덴마크, 네덜란드 및 주요 빅테크 기업의 사례를 각각 살펴 본다.

---

52) UNITED NATIONS(2021), para.37-38.
53) Ad Hoc Committee on Artificial Intelligence Policy Development Group(2021), p.22.
54) The Danish Institute for Human Rights(2020b). Introduction, Chapter 2.5.2.
55) 김동현(2022), 135면.

제1절 인공지능 위험영향평가 사례

1. 유럽연합 위험영향평가

유럽연합은 신기술에 대하여 위험기반접근법을 취해 왔고, 여러 법률에서 신기술을 도입하는 기관 및 기업에 위험평가의 실시를 요구하여 왔다. 2016년 제정된 <일반개인정보보호규정>은 자동화된 개인정보처리를 비롯한 고위험 개인정보처리에 대하여 '개인정보보호 영향평가'를 적용하여 왔다. 인공지능의 발달로 안전 및 인권에 대한 침해 우려가 커짐에 따라 유럽연합 집행위원회는 2019년 <신뢰할 수 있는 인공지능 평가 목록>을 제안하고 2021년 제안한 <인공지능법(안)>에서 고위험 인공지능에 대한 위험평가를 제도화하였다. 한편, 2022년 11월 16일 발효한 유럽연합 <디지털서비스법>은 체계적 위험이 우려되는 대규모온라인플랫폼의 알고리즘 등에 대하여 위험평가를 실시하도록 하였다.

가. 자동화된 의사결정과 DPIA

2016년 제정되어 2018년 시행된 유럽연합 <일반개인정보보호규정(General Data Protection Regulation, 이하 'GDPR')>은 자연인의 자유와 권리에 '고위험'(high risk)을 야기할 수 있는 개인정보처리에 대하여 공공과 민간을 불문하고 개인정보처리자(controller)가 '개인정보보호 영향평가(Data Protection Impact Assessment, 이하 'DPIA')'를 실시하여 해당 위험을 평가하고 이를 해결할 수 있는 조치를 결정하여 위험을 관리하도록 하였다.

GDPR은 DPIA를 실시하여야 하는 고위험 개인정보처리에 대하여 △프로파일링 등의 자동화된 처리에 기반한, 개인에 관한 개인적 측면을 체계적이고 광범위하게 평가하는 것으로 해당 평가에 근거한 결정이 해당 개인에게 법적 효력을 미치거나 이와 유사하게 개인에게 중대한 영향을 미치는 경우, △민감정보에 대한 대규모 처리나 범죄경력 및 범죄 행위에 관련된 개인정보에 대한 처리, △공개적으로 접근 가능한 지역에 대한 대규모의 체계적인 감시가 포함된다고 규정하였다(제35조 제3항). GDPR은 특히 '신기술'을 사용하는 처리 유형이 개인의 권리와 자유에 중대한 위험을 초래할 것으로 예상되는 경

우 DPIA의 실시가 필요하다고 명시하였다(제35조 제1항 및 전문 91). 이를 보다 구체화한 DPIA 가이드라인[56]은 고위험 개인정보처리를 9가지로 설명하였다.

유럽연합 GDPR 고위험 개인정보처리

① 평가나 점수화. 특히 "정보주체의 작업 역량, 경제적 상황, 건강, 개인적 선호나 기호, 신뢰성이나 행동, 위치나 이동과 관련된 측면"에 대한 프로파일링 및 예측의 경우. 예를 들어, 신용평가, 질병 예측을 위한 유전자 검사, 이용기반 마케팅 프로파일링 구축 등

② 법적 혹은 이와 유사한 중대한 효과를 미치는 자동화된 의사결정

③ 체계적인 감시. 정보주체가 인지하지 못하는 사이에 개인정보가 수집, 이용될 수 있는 경우. 특히 개인들이 처리의 대상이 되는 것을 피하기 어려운 공공장소에서 이루어지는 체계적인 감시가 이에 해당함

④ 민감정보 및 범죄경력정보 또는 매우 사적인 개인정보의 처리. 예를 들어, 일반 병원 또는 사설탐정이 처리하는 개인정보처리나 통신비밀, 위치정보, 금융정보 등

⑤ 대규모로 처리되는 개인정보. 대규모 처리는 △정보주체의 수(절대수 및 인구 비율) △처리되는 개인정보의 양과 범위 △개인정보처리 행위의 지속성, 영구성 △처리행위의 지리적 범위를 고려함

⑥ 데이터셋이 연계 및 결합되는 경우. 정보주체의 합리적 기대를 벗어나 다른 개인정보처리자에 의해, 다른 목적을 위해 처리되는 둘 이상의 데이터셋 처리가 해당함

⑦ 취약한 정보주체에 대한 개인정보처리. 아동, 피고용인, 지적장애인, 망명인, 노인, 환자 등 취약한 정보주체의 개인정보를 처리할 때 개인정보처리자와 정보주체의 불균등한 권력관계를 고려하여야 함

⑧ 혁신적인 사용 또는 새로운 기술적, 조직적 해결책의 적용. 예를 들어 물리적 접근통제를 위해 지문이나 얼굴인식기술을 사용하는 경우. 새로운 형식의 처리가 이루어질 수 있고 그 개인적, 사회적 결과가 알려지지 않은 경우. 예를 들어, 새로운 형식의 IoT 기기의 처리가 이에 해당함

56) Article 29 Working Party(2017). Guidelines on Data Protection Impact Assessment(DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, wp248rev.01. <https://ec.europa.eu/newsroom/article29/items/611236(접근일: 2022. 9. 15.)>;

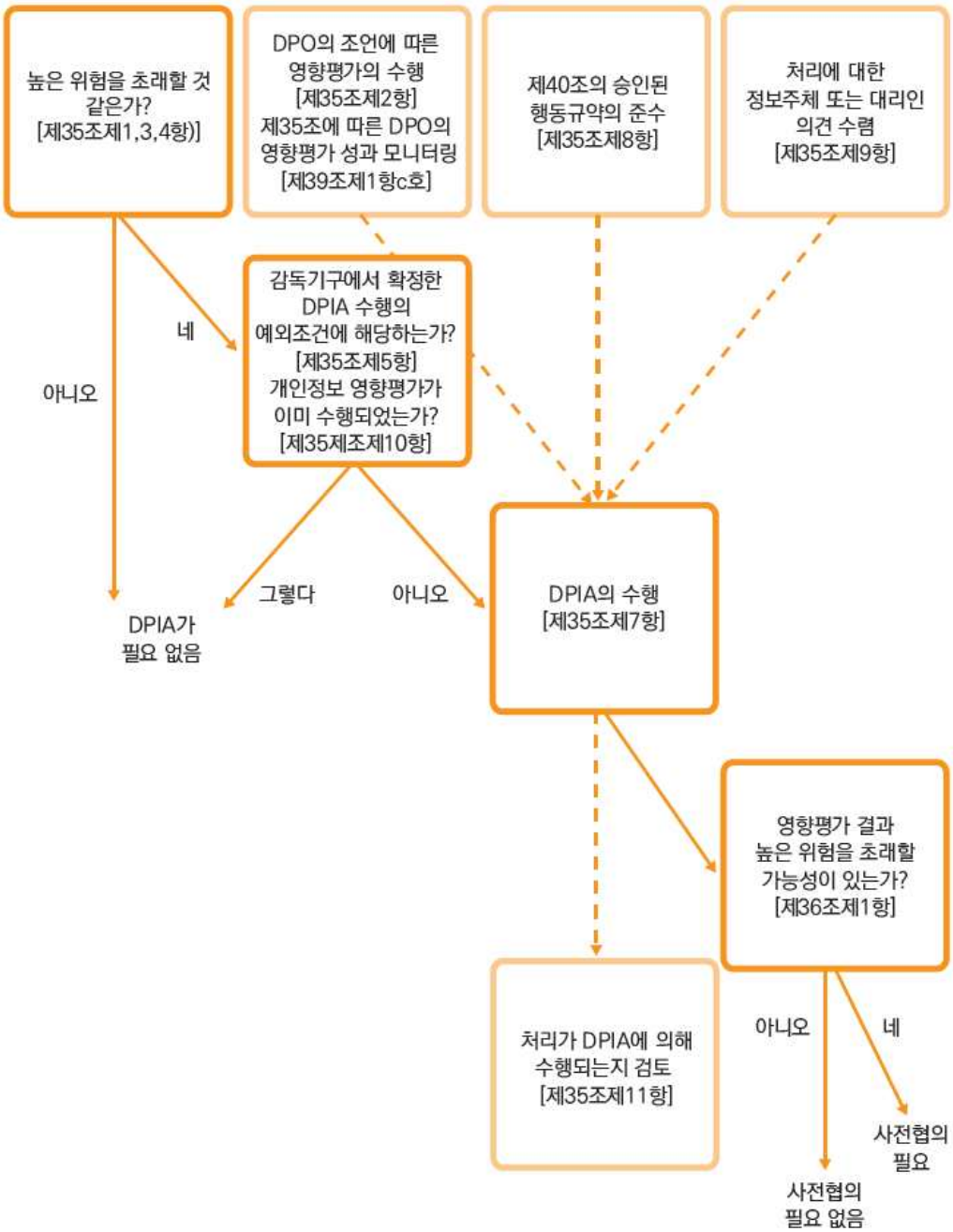
⑨ 개인정보처리 자체가 정보주체가 자신의 권리를 행사하거나 서비스를 받거나 계약 체결을 방해할 경우. 즉, 정보주체의 서비스 접근이나 계약 체결을 허용, 수정, 거부하는 것을 목적으로 하는 개인정보처리. 예를 들어, 은행에서 고객의 신용도를 평가하는 경우가 이에 해당함

인공지능 기술의 발전으로 프로파일링과 자동화된 의사결정을 비롯한 고위험 개인정보처리가 인공지능에 기반하여 이루어지는 경우가 많아졌다. 개인정보처리자는 고위험 개인정보처리를 실시하기 전에 위험요소의 출처, 성격, 특성 및 심각성이 GDPR 준수에 위배되지 않는지 평가하는 DPIA를 실시하여야 한다. DPIA의 실시 후 개인정보처리로 초래되는 위험에 변화가 생긴 경우에도 이 처리가 부합하는지 검토하여야 한다.

개인정보처리자는 해당 고위험을 완화하고 개인정보보호를 보장하며 GDPR을 준수했음을 입증하는데 예상되는 조치, 안전조치 및 메커니즘을 DPIA 절차에 포함해야 한다. 또한 개인정보처리자는 예정된 처리에 대하여 개인정보주체 또는 그 대리인의 의견을 구해야 한다.

만약 완화할 수 없는 고위험이 있다면 개인정보처리자는 개인정보보호 감독기구와 의무적으로 협의하여야 한다. DPIA의 구체적인 사항은 개인정보처리자의 재량에 속하지만, 고위험 개인정보처리에 대한 평가 의무 자체와 감독기구에 대한 협의 의무를 위반할 경우 과징금이 부과된다.

[그림 6] 유럽연합 DPIA 수행 단계[57]



57) 한국인터넷진흥원(2018). 우리 기업을 위한 'EU 일반 개인정보 보호법(GDPR)' 가이드북, 121면.

나. 고위험 인공지능과 위험평가

유럽연합 집행위원회 인공지능 고위전문가그룹은 2019년 <신뢰할 수 있는 인공지능 윤리 지침>[58]을 마련하였다. 이 지침은 인공지능 시스템이 윤리 지침의 요구사항을 만족하는지 평가하기 위하여 <신뢰할 수 있는 인공지능 평가 목록(Assessment List for Trustworthy Artificial Intelligence, 이하 'ALTAI 평가 목록')>의 초안을 공개하였다. 고위전문가그룹은 평가 목록의 초안에 대한 이해관계자들의 의견을 수렴하고 파일럿테스트를 거쳐 2020년 7월 ALTAI 최종 평가 목록을 발표하였다.[59]

ALTAI 평가 목록은 윤리 지침의 요구사항이 인공지능 개발 수명주기에서 충족되는지를 실무적으로 점검할 수 있도록 구성되어 있으며, 인공지능 영향평가의 초기 형태로 알려졌다. 유럽연합 인공지능 윤리 지침은 이 평가 목록과 더불어 선언적인 수준에 그쳤던 여타의 인공지능 윤리 가이드라인들과 다른 실천력을 보여주었다.

구체적으로 ALTAI 평가 목록은 ①인간행위자와 감독, ②기술적 견고성과 안정성, ③프라이버시 및 데이터 거버넌스, ④투명성, ⑤다양성, 차별금지, 공정성, ⑥사회·환경적 복지, ⑦책무성 등 윤리지침의 7대 요구사항의 준수 여부를 점검할 수 있는 140여 개 객관식·주관식 질의 문항들로 구성되어 있다.

특히 ALTAI는 윤리지침의 요구사항을 점검하기에 앞서 '기본권 영향평가(fundamental rights impact assessment)'를 수행할 것을 제안하였다. 이 기본권 영향평가는 유럽인권조약 및 의정서, 유럽사회헌장 조항을 인용하며 그 준수에 대하여 질의하는 12개 문항으로 구성된다.

---

58) High-Level Expert Group on Artificial Intelligence(2019).
59) High-Level Expert Group on Artificial Intelligence(2020). 평가 목록의 질의 문항 번역은 부록 I 참조.

[표 3] 유럽연합 ALTAI 평가 목록[60]

카테고리(문항 수)		평가 내용
기본권 보장 (12)		인간의 기본권에 대한 영향평가를 수행하고, 상충되는 원칙과 권리 간 절충 관계를 확인하고 문서화함
인간행위자와 감독	인간행위자와 자율성 (14)	의사결정의 자동화 수준을 고려할 때 사람과의 상호작용과 통제・감독 수준이 적절한지 확인
	인간의 감독 (8)	Human-in-the-loop, Human-on-the-loop, Human-in-command와 같은 시스템 통제 메커니즘의 확보 여부, 필요할 때 안전한 중단 수단・절차를 구비했는지 확인
기술적 견고성과 안정성	공격에 대한 회복성과 보안 (9)	시스템의 취약성과 잠재적 위험을 검토하고 대비하여 회복성과 안전성을 보장하기 위한 대비 계획・절차・보험・거버넌스 등 구비 여부 확인
	일반적 안전성 (10)	위험 정의, 위험 지표 및 수준 관리, 잠재적 위협 식별 및 예상 결과 검토 여부, 핵심 시스템의 의존성 등 평가 방법 확보 여부
	정확성 (5)	데이터의 품질, 시스템의 최신성, 정확성 확보 노력 등 평가
	신뢰성, 대체계획 및 재현성 (9)	신뢰성과 재현성의 중요성, 신뢰성 및 재현성의 보장 방법, 시스템 에러에 대비한 대체 계획의 여부
프라이버시 및 데이터 거버넌스	프라이버시 (2)	민감한 데이터를 최소화하고 적절한 개인정보보호 강화조치 여부 확인
	데이터 거버넌스 (11)	데이터의 품질 및 무결성 감독 절차의 수립 여부, 국제 관리 표준의 준수 여부 확인
투명성	추적가능성 (6)	시스템을 설계・개발・테스트・검증하는데 사용되는 방법과 결과 활용에 대한 의사결정 사항들을 문서화하여 추적이 가능한지 확인함
	설명가능성 (2)	시스템에 의한 선택이나 자동화된 의사결정이 설명이 가능한지 확인함

60) 유재흥 외(2021). 유럽(EU)의 인공지능 윤리 정책 현황과 시사점: 원칙에서 실천으로. 소프트웨어정책연구소 ISSUE REPORT IS-114(2021. 3. 25), 14면, 일부수정.

	고지 (5)	시스템의 목적・용례・특성・한계를 이해관계자에게 충분히 고지했는지 확인함
다양성, 차별금지, 공정성	불공정한 편향 회피 (15)	시스템 설계 시 사용한 공정성의 정의가 목적, 사용자 등에 적절한 수준인지, 불공정한 편향을 방지하기 위한 절차와 메커니즘을 수립하고 투명하게 고지했는지 확인함
	접근가능성 및 보편적 설계 (9)	접근성・이해관계자의 다양성을 보장하고 보편적인 설계를 채택했는지 확인함
	이해관계자 참여 (1)	시스템의 설계 및 개발 과정에서 다양한 이해관계자의 참여 기회 여부
사회・환경적 복지	환경적 복지 (4)	지속가능하고 환경친화적인 방식의 시스템 구현 및 운영 여부
	일자리 및 역량에 대한 영향 (8)	시스템이 노동 환경, 고용 관계, 직무 등에 미치는 영향을 평가
	사회 전반 및 민주주의에 대한 영향 (4)	제도, 민주주의, 거시적 차원에서의 시스템의 사회적 영향력 평가
책무성	감사가능성 (2)	시스템의 절차 및 결과물에 대한 감사 기능을 적절하게 구현하여 추적성과 감사가능성을 보장하고 있는지, 상충되는 원칙 간 절충에 대한 의사결정이 문서화 되었는지 확인함
	위험관리 (10)	시스템의 영향평가와 책임성 강화 교육・훈련을 하고 있는지, 인공지능 윤리검토위원회를 설치하고 내부 지침을 마련했는지, 제3자나 노동자에게 위험보고 절차를 수립했는지 확인함

이 윤리 지침은 발표 시점부터 “새롭고 구체적인 규제 마련을 장려”[61]하려는 지향점을 가지고 있었으며, 유럽연합 집행위원회는 2019년 4월 위 윤리 지침을 공식 채택하고 이후 윤리 지침을 실현하는 인공지능 규제 프레임워크(regulation framework)를 마련해 왔다. 2020년 2월 발표된 유럽연합 <인공지능 백서>[62]는 고위험 인공지능을 중심으로 그 위험을 규제하는 프레임워크를 제시하였다.

61) High-Level Expert Group on Artificial Intelligence(2019), 11p.
62) European Commission(2020a).

유럽연합은 2020년 5월 발표한 <인공지능 공공조달 백서>[63]에서 국민의 기본권에 영향을 미치는 공공부문 인공지능 조달에 있어 위험기반·체계적 접근법을 취하고, 이를 위한 5단계 실사 절차를 권장하였다. 첫째, '위험영향평가'를 사전적으로 실시하여 사람과 집단, 권리와 자유, 민주적 조직과 절차, 사회와 환경에 부작용을 미치는지를 살피도록 하였다. 둘째, 공급자 예비 심사를 통하여 설계 절차의 최초 단계서부터 인공지능 관련 데이터 윤리 요구사항을 고려하고 정의하고 구현하도록 요구하였다. 이때 요구되는 데이터 윤리 요구사항으로는 △인공지능이 챗봇, 가상비서 등으로 이용자와 직접적으로 상호작용하거나 자동화된 의사결정 등으로 간접적으로 상호작용한다면 필히 인간이 아니라는 점을 밝혀야 하고, △인공지능 시스템이 추적가능하고, 설명가능하고 이해관계자를 수용해야 하며, △인공지능 시스템은 편향을 방지하고 보편적 설계를 따라야 하며 검토 절차를 포함해야 하고, △기술적 안전성이 문서화되어 설명가능성, 공정 커뮤니케이션 및 감사를 보장해야 한다. 셋째, 공급자를 선정하는 품질 기준에 정보보안, 데이터 윤리, 환경, 사생활, 보편적 설계 등에 적용되는 표준 및 시스템 기술사양을 반영해야 한다. 넷째, 발주 공공기관은 계약 조건에 지속가능성, 기본권 존중, 데이터 윤리에 대한 조항을 포함하고 제재 조항 및 문서화 요구사항을 명시해야 한다. 다섯째, 공급자는 계약을 집행할 때 데이터 윤리, 법적 준수, 책임, 기술적 안전성, 지속가능성 등 요구사항을 충족해야 한다.

한편, 유럽연합 집행위원회는 2021년 4월 21일 제안한 인공지능법(안)(Artificial Intelligence Act)[64]에서 앞서 윤리 지침의 원칙들을 법규로 발전시켰다. 인공지능법(안)은 고위험 인공지능 시스템에 대하여 위험관리, 데이터관리, 품질 관리, 기술문서 작성 등을 의무화하고 이를 기반으로 출시 전 적합성 평가와 출시 후 모니터링을 요구하였다. 고위험 인공지능 시스템의 유통시에는 규정 준수를 나타내는 CE 적합성 인증마크를 표시해야 한다.

고위험에 속하는 인공지능 시스템으로는 항공, 자동차, 철도, 기계, 장난감, 승강기, 의료기기 등 안전 제품 인공지능(부속서Ⅱ)과 생체인식, 중요인프라·응급, 교육평가·입시, 고용·직원관리, 사회복지급여·신용평가, 법집행, 출입국관리, 사법업무 등 기본권에 영

63) European Commission(2020b).
64) European Commission(2021).

향을 미칠 수 있는 독립형 인공지능 시스템(부속서Ⅲ)이 있다. 반면 인간이 인지하지 못하는 방식으로 행동, 의견 또는 결정을 조작하여 신체적·정신적 위험을 가져올 수 있는 인공지능, 아동·장애인 등의 취약성 또는 특수 상황을 표적으로 삼는 인공지능, 공공기관이 자연인의 신뢰도를 예측하거나 특성을 분류하여 불리한 대우를 하는 사회신용평가 인공지능, 법집행기관의 실시간 공공장소 원격 생체인식 인공지능은 원칙적으로 금지된다.

고위험 인공지능 시스템에 대한 위험관리 시스템의 경우, 고위험 인공지능 시스템의 수명주기 전반에 걸쳐 지속적으로 운영되고 정기적·체계적으로 갱신을 반복해야 하며, (a) 고위험 인공지능 시스템과 관련된 알려지고 예측가능한 위험의 식별 및 분석, (b) 고위험 인공지능 시스템을 합리적으로 예측 가능한 오남용 조건 하에서 원래 목적으로 사용할 때 발생할 수 있는 위험의 추정 및 평가, (c) 출시 후 모니터링 시스템에서 수집한 데이터의 분석에 근거한 기타 위험 발생가능성의 평가, (d) 적합한 위험 관리 수단의 채택으로 구성된다(제9조 제2항). 특히 설계와 개발을 통해 최대한 위험을 제거 또는 완화하고, 제거할 수 없는 위험에 대해서는 완화 및 통제 조치를 시행하는 등 적합한 위험 관리 수단을 채택한 후에도 잔여 위험이 남아 있을 경우, 이를 허용 가능한 수준으로 보장하고 사용자에게 통지해야 한다(동조 제4항). 고위험 인공지능 시스템은 테스트를 통해 원래 목적에 일치하고 요구사항을 준수하는지 여부를 확인해야 하며(동조 제5항), 이러한 테스트는 개발 과정에서 임의의 시점에 수행하되, 어떠한 경우에도 출시 또는 서비스 개시 전에 수행되어야 한다(동조 제7항). 특히 아동이 고위험 시스템에 접근하거나 영향을 받을 가능성이 있는지 여부에 각별한 주의를 기울여야 한다(동조 제8항).

다. 대규모온라인플랫폼과 위험평가

2000년 유럽연합 전자상거래 지침(Directive 2000/31/EC)을 대체하는 규정(Regulation)으로 2022년 11월 16일자로 발효한 유럽연합 디지털서비스법(Digital Service Act)[65]은 대규모온라인플랫폼에 대한 위험평가를 의무화하였다. 대규모온라인플랫폼은 역내에서 유럽

65) Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

연합 전체 인구의 10%에 해당하는 월평균 4,500만명 이상 활성 이용자에게 최소 4개월 간 연속적으로 서비스를 제공한 온라인플랫폼을 말한다.

유럽연합은 이들 대규모 온라인 플랫폼에 다음과 같은 체계적인 위험이 있다고 지적한다. 첫째, 불법적 서비스 오남용의 위험이 있다. 아동성폭력 콘텐츠, 불법 혐오표현 유포, 위조품 등 법률에서 금지하는 상품과 서비스 제공 등이 이에 해당한다. 둘째, 기본권에 미치는 부정적 영향의 위험이 있다. 인간 존엄성, 사생활권과 개인정보에 대한 권리, 표현의 자유와 정보의 자유, 언론의 자유와 다양성, 차별받지 않을 권리, 아동 권리와 소비자 권리 등에 대한 위험이 이에 해당한다. 셋째, 시민 담론, 선거 과정 및 공공 안전에 미치는 실제적 또는 잠재적 영향의 위험이 있다. 넷째, 젠더 폭력, 공중 보건, 아동 보호 및 개인의 신체적, 정신적 건강에 미치는 실제적이고 예상되는 부정적인 영향의 위협이 있다.

이에 대규모온라인플랫폼은 연 1회 이상 또는 중대한 위험을 미치는 기능을 배치하기 전에, 알고리즘시스템 등에 대한 설계, 서비스 기능 및 사용에서 기인하는 체계적인 위험을 식별, 분석, 평가하는 위험평가를 실시하여야 한다. 위험평가는 체계적인 위험의 심각도(severity) 및 발생가능성을 고려하여 비례적이고 해당 서비스에 특화되어 있어야 한다.

위험평가를 수행할 때에는 특히 △추천시스템 및 관련 알고리즘시스템의 설계 △콘텐츠 관리시스템 △이용약관 및 그 집행 △광고 선택 및 표시 시스템 △플랫폼 제공자의 관행 관련 데이터들이 체계적인 위험에 영향을 미치거나 어떻게 미치는지를 검토하여야 하고, 이용약관에 위배되는 불법 콘텐츠의 확산 및 고의적인 서비스 조작에 대하여 분석하고 특정 지역적·언어적 측면을 고려하여야 한다. 평가 문서는 평가 후 3년간 보관하며 관할기관 요청시 제출하여야 한다(제34조).

대규모온라인플랫폼은 식별된 체계적 위험에 대하여 합리적이고 비례적이며 효과적인 위험 완화조치를 취하면서 이들 조치가 기본권에 미치는 영향도 고려하여야 한다. 완화조치에는 다음과 같은 조치가 포함된다. (a) 온라인 인터페이스를 비롯한 서비스 설계, 특성 및 기능에 대한 변경, (b) 이용약관과 집행에 대한 변경, (c) 불법적 혐오 발언 및 사이버 폭력 등 불법 콘텐츠에 대한 신속한 삭제나 접근차단 등 콘텐츠 관리 및 관련 의사결정 절차에 대한 변경 및 자원 할당, (d) 추천시스템 등 알고리즘시스템의 테스트

및 변경, (e) 광고시스템의 변경 및 광고표시를 제한하거나 조정하는 조치, (f) 체계적인 위험 탐지와 관련한 업무에 대한 내부 절차, 자원, 테스트, 문서화 및 감독의 강화, (g) 신뢰할 수 있는 신고자(trusted flaggers)와 협력·조정 및 법정외 분쟁해결기구의 결정 이행, (h) 행동강령 및 위기프로토콜을 통해 다른 플랫폼과 협력 또는 조정, (i) 이용자 정보 제공 개선을 위한 인식 확산 조치 및 온라인 인터페이스 변경, (j) 연령 확인 및 부모 통제 도구, 아동 폭력에 대해 신고하거나 지원하는 도구 등 아동 권리 보호를 위한 조치, (k) 실존하는 사람, 물체, 장소 및 기타 실체물 또는 사건에 대한 딥페이크 이미지, 오디오 또는 비디오가 온라인 인터페이스에 표시될 때 명확한 표시로 구별하는 조치 등(제35조 제1항).

유럽연합 디지털 서비스 이사회(European Board for Digital Services)는 대규모온라인 플랫폼이 보고한 위험평가 및 기타 정보를 종합하여 가장 두드러지고 반복적인 체계적 위험을 식별하고 평가한 내용을 연간보고서로 공개하고 완화조치에 대한 모범사례를 제시한다. 집행위원회는 공개적인 협의를 거쳐 특정 위험에 대한 조치를 권고하는 지침을 발행할 수 있다(제35조 제2항 및 제3항). 한편 관할기관의 지정을 받은 학술연구자들(vetted researcher)은 대규모온라인플랫폼의 데이터에 접근하여 체계적 위험에 대해 탐지·식별·이해하고, 위험 완화 조치의 적절성·효율성·영향을 평가하는 데 기여하는 연구를 수행할 수 있다(제40조 제4항). 이들 연구자는 상업적인 이익으로부터 독립적이어야 하기 때문에 신청시 연구자금 출처를 공개하고 연구 완료 후 연구 결과 역시 무료로 공개한다.

2. 캐나다 알고리즘영향평가

캐나다 정부는 2019년 연방정부의 조달 정책 및 제도를 소관하는 재정위원회의 훈령으로 「자동화된 의사결정 훈령(Directive on Automated Decision-Making)」을 제정하여 시행 중이다.[66]

66) Government of Canada. Directive on Automated Decision-Making. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592(접근일: 2022. 8. 15.)>; 김기중 외(2021), 85-86면.

이 훈령은 공공기관이 의사결정에 사용하는 인공지능 시스템에 대하여 위험기반 접근법을 취하고 있으며, 자동화된 의사결정이 개인·공동체의 권리, 개인·공동체의 건강 및 복리, 개인·단체·공동체의 경제적 이익, 생태계의 지속가능성에 미치는 영향은 물론 그 영향의 지속성과 회복가능성에 따라 그 위험 수준을 4단계로 구분하였다.

특히 훈령 제6장은 인공지능 시스템에 대한 요구사항(Requirements)을 법규화하고 있는데, 평가된 위험 수준이 높을수록 적용되는 요구사항도 높아진다. 이 요구사항은 △동료 검토(peer review), △고지(notice), △의사결정에 대한 인간의 개입(Human-in-the-loop for decisions), △설명 요구사항(Explanation Requirement), △교육훈련(Training), △비상계획(Contingency Planning), △시스템 구동 승인(Approval for the system to operate)에 대한 것이다. 예를 들어 직원의 교육훈련에 대한 요구사항의 경우, 위험이 낮은 수준 1의 시스템에는 적용되지 않지만, 수준 2 위험의 시스템은 시스템의 설계 및 기능에 대하여 문서화하도록 하였고, 수준 3 위험의 시스템은 문서화 외에도 교육 과정을 필수적으로 이수하도록 하였다. 가장 위험도가 높은 수준 4의 시스템은 문서화, 교육과정 이수에 더하여 이를 반복 이수하도록 하고 이수 사실을 객관적으로 확인할 수 있는 방법을 마련하도록 하였다.

가. 평가의 절차

캐나다 정부는 요구사항의 기반이 되는 위험 수준을 측정하기 위하여 알고리즘영향평가도구(Algorithmic Impact Assessment, 이하 ‘캐나다 알고리즘영향평가’)를 개발하였다. 의사결정에 알고리즘을 사용하는 공공기관은 평가를 의무적으로 실시하여야 한다.

실시 시기는 프로젝트 설계 단계 초기에 우선 실시하고, 시스템의 생산 전에도 두 번째로 실시하여 요구사항이 구축된 시스템에 반영되었는지 확인하도록 하였다. 두 번째 평가 결과는 일반 접근이 가능한 형식으로 온라인에 공개하여야 한다. 시스템의 기능 또는 범위가 변경되면 평가를 갱신하여야 한다.

캐나다 재정위원회 자동화된 의사결정 훈령

6.1. 알고리즘영향평가

6.1.1. 자동화된 의사결정 시스템을 생산하기 전에 알고리즘영향평가를 완료한다.

6.1.2. 알고리즘영향평가에 의해 판단이 내려진 경우 부록 C에 규정된 관련 요구사항을 적용한다.

6.1.3. 자동화된 의사결정 시스템의 기능 또는 범위가 변경된 경우 알고리즘영향평가를 갱신한다.

6.1.4. 알고리즘영향평가의 최종 결과를 <열린 정부 훈령>에 부합하도록 캐나다 정부 웹사이트 및 캐나다 재정위원회가 지정한 기타 서비스를 통해 일반 접근이 가능한 형식으로 공개한다.

나. 평가의 기준

평가도구인 캐나다 알고리즘영향평가는 자동화된 의사결정 시스템의 위험 및 완화 정도에 대하여 묻고 영향 수준을 판단하는 질문지로 구성되어 있다.[67] 위험 점수는 48개의 위험과 33개의 완화 조치에 대한 질의와 그 답변에 기반하여 산출되며, 완화 점수가 80% 이상 도달하면 위험 점수를 15% 차감한다.

위험을 측정하는 질의는 프로젝트, 시스템, 알고리즘, 의사결정, 영향, 데이터에 대한 다방면 질의로 이루어지며, 완화에 대한 질의는 이해관계자 협의, 데이터 품질, 절차적 공정성, 개인정보보호에 대한 내용들이다.

67) Government of Canada. Algorithmic Impact Assessment Tool. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html(접근일: 2022. 8. 15).

캐나다 알고리즘영향평가 완화 조치 관련 질의[68]

제10장 : 자문

다음 집단이 참여합니까?

내부 이해관계자(전략 정책 및 계획 부서, 데이터 거버넌스 부서, 프로그램 정책 부서 등) [1점]
☐ 예
☐ 아니오

외부 이해관계자(시민사회, 학계, 산업계 등) [1점]
☐ 예
☐ 아니오

제11장 : 위험성 제거 및 완화 조치 - 데이터 품질

편향성 및 기타 예상치 못한 결과물에 대해 데이터셋을 검사할 때 문서화된 절차를 이용합니까? 이러한 절차에는 프레임워크, 방법론, 지침 또는 기타 평가도구를 적용하는 활동 등이 포함됩니다. [3점]
☐ 예
☐ 아니오

설계 단계에서 데이터 품질 문제를 어떻게 해결하였는지 문서화하는 절차를 둡니까? [1점]
☐ 예
☐ 아니오

위 정보가 공개됩니까? [1점]
☐ 예
☐ 아니오

이 데이터에 대하여 '젠더 기반 분석 플러스'를 적용합니까? [1점]
☐ 예
☐ 아니오

위 정보가 공개됩니까? [1점]
☐ 예
☐ 아니오

이 시스템의 설계, 개발, 유지보수 및 개선에 대한 책임을 소속 기관 내에서 할당했습니까? [2점]
☐ 예
☐ 아니오

오래된 데이터나 신뢰할 수 없는 데이터가 자동화된 의사결정에 사용되는 위험을 관리하기 위하여 문서화된 절차를 두고 있습니까? [2점]
☐ 예
☐ 아니오

위 정보가 공개됩니까? [1점]
☐ 예
☐ 아니오

이 시스템에 사용된 데이터가 정보공개포털에 공개됩니까? [2점]
☐ 예
☐ 아니오

제12장 : 위험성 제거 및 완화 조치 - 절차적 공정성

감사추적(audit trail) 기능이 법률에서 명시한 권한 또는 위임된 권한을 확인합니까? [1점]
☐ 예
☐ 아니오

이 시스템은 시스템에 의해 수행된 모든 권고사항 또는 의사결정을 기록하는 감사추적 기능을 제공합니까? [2점]
☐ 예
☐ 아니오

모든 주요 의사결정 포인트가 감사추적 정보에서 파악됩니까? [2점]
☐ 예
☐ 아니오

자동화된 시스템 로직 내의 모든 주요 의사결정 포인트들이 관련 법률, 정책 또는 절차에 링크되어 있습니까? [1점]
☐ 예
☐ 아니오

모델과 시스템의 모든 변경 사항을 상세히 기록한 로그기록을 관리합니까? [2점]
☐ 예
☐ 아니오

감사추적 기능이 시스템에 의해 이루어진 모든 결정 포인트를 명확하게 설명합니까? [1점]

☐ 예
☐ 아니오

시스템이 생성한 감사추적 기능을 사용하여 필요한 경우 결정 통지(이유서 또는 기타 통지 포함)를 생성할 수 있습니까? [1점]

☐ 예
☐ 아니오

감사추적 기능이 시스템의 각 결정에 어떤 버전의 시스템이 사용되었는지 정확하게 파악할 수 있습니까? [2점]

☐ 예
☐ 아니오

감사추적 기능이 해당 의사결정에 대해 권한을 가진 사람을 파악할 수 있습니까? [1점]

☐ 예
☐ 아니오

시스템이 필요할 경우 의사결정이나 권고사항에 대한 이유를 제시할 수 있습니까? [2점]

☐ 예
☐ 아니오

시스템에 대한 접근 권한을 부여, 모니터링 및 취소하는 절차가 있습니까? [1점]

☐ 예
☐ 아니오

시스템 사용자가 피드백을 포착할 수 있는 메커니즘이 있습니까? [1점]

☐ 예
☐ 아니오

의사결정에 대해 이의를 제기하고자 하는 당사자를 위해 계획 또는 수립된 소구 절차가 있습니까? [2점]

☐ 예
☐ 아니오

시스템의 결정에 대해 인간의 기각이 가능합니까? [2점]

☐ 예
☐ 아니오

기각이 이루어졌을 때 그 사실을 기록하는 절차가 있습니까? [1점]
☐ 예
☐ 아니오

감사추적 기능에 시스템의 작동 또는 수행에 대한 수정사항을 기록하는 변경사항 관리 절차가 포함됩니까? [2점]
☐ 예
☐ 아니오
캐나다 정부 정보기술아키텍처 검토위원회(Government of Canada Enterprise Architecture Review Board)에 컨셉 케이스를 제출할 예정입니까? [1점]
☐ 예
☐ 아니오

제13장 : 위험성 제거 및 완화 조치 - 개인정보보호

시스템에 개인정보 사용이 포함된 경우, 개인정보보호 영향평가를 수행하였거나 수행할 예정이거나, 기존 영향평가를 갱신할 예정입니까? [1점]
☐ 예
☐ 아니오

프로젝트의 개념 수립 단계에서부터 시스템에 보안과 개인정보보호조치를 설계하고 구축합니까? [1점]
☐ 예
☐ 아니오

개인정보가 폐쇄형 시스템(예: 인터넷, 인트라넷, 기타 시스템에 연결하지 않음) 내에서 사용됩니까? [1점]
☐ 예
☐ 아니오

개인정보 공유와 관련된 경우, 적절한 보호조치가 수반된 동의서 또는 약정이 수립되어 있습니까? [1점]
☐ 예
☐ 아니오

68) Government of Canada. “Algorithmic Impact Assessment v0.9.1”. https://open.canada.ca/aia-eia-js/?lang=en(접근일: 2022. 8. 15.); 김기중 외(2021), 부록 V, 일부 수정.

다. 다른 규제 메커니즘과의 관계

훈령은 자동화된 의사결정 시스템의 책임 있는 사용을 지원하기 위해 정부 공통의 전략, 접근법 및 절차를 개발하고 범정부 차원에서 또는 타부문 관할 기관과 교류하고 참여하도록 하였다(훈령 8.4). 예를 들어 평가 중인 알고리즘 시스템이 개인정보와 관련된 경우, 캐나다 알고리즘영향평가는 “개인정보보호 영향평가를 수행하거나 수행한 적이 있거나, 기존 영향평가를 갱신할 예정입니까?” 고 묻는 등 개인정보 보호법의 원칙과 규정을 참조하였다.

라. 운영상 쟁점

1) 평가 수행 주체

이 훈령은 외부 대상으로 추천이나 행정적 의사결정을 하기 위하여 알고리즘 시스템, 도구, 통계적 모델을 개발하거나 조달하는 연방정부 기관(국가정보 시스템 등 일부 시스템은 제외)을 그 적용 범위로 한다(훈령 5). 평가는 자동화된 의사결정 시스템을 생산하려는 해당 기관이 온라인에서 직접 수행한다.

이 훈령은 캐나다 정부 재정위원회가 소관하지만, 평가 기준이나 점수 산정 등 알고리즘영향평가의 개발 및 유지관리는 캐나다 최고 정보 책임자(CIO)가 담당한다(훈령 8.3). 훈령을 준수하지 않는 경우 재정위원회의 조치 대상이 될 수 있다(훈령 7.1).

2) 이해관계자 참여

캐나다 알고리즘영향평가는 완화 조치가 이루어지면 위험 점수를 차감하고 그에 따른 요구사항도 낮추는 방식으로 각 기관이 자발적이고 능동적으로 완화 조치를 취하도록 유도하고 있다. 이해관계자 참여는 이 완화 조치의 주요 질의 항목에 해당한다.

캐나다 알고리즘영향평가는 내부 이해관계자와 외부 이해관계자의 참여에 대하여 구분하여 각각 질의한다. 내부 이해관계자로는 전략적 정책 및 계획 부서, 데이터 거버넌스 부서, 프로그램 정책 부서, 법률 자문, 정보공개 및 개인정보보호 담당, 소통 담당, 고

객 관리 부서 등을 포함하였고, 외부 이해관계자로는 시민사회, 학계, 산업계, 개인정보 보호 감독기구 등을 포함하였다.

3) 평가 결과 공개

훈령은 평가의 최종 결과를 캐나다 정부 웹사이트 및 캐나다 재정위원회가 지정한 서비스에 일반 접근이 가능한 형식으로 공개하도록 하였다(훈령 6.1.4.).

이에 캐나다 알고리즘영향평가 정보공개 페이지에는 연방공공보건청의 코로나19 예방접종증명서 인식사업(ArriveCAN) 등 공공기관 알고리즘 시스템에 대한 평가 결과가 공개되어 있다[69]. ArriveCAN 시스템의 경우 OCR 기술을 사용하여 캐나다 입국자의 예방접종증명서를 입력받고 출입국관리업무의 입국자격심사나 격리 결정을 지원하는 시스템이다. 이 시스템은 입국자의 개인정보를 이미지 형태로 수집하고, 개인의 사생활의 권리, 이동의 자유, 건강권 등에 관여하며, 알고리즘에 대한 영업비밀 보호, 알고리즘 개발기관 외부 이용, 네트워크 연결 등이 이루어진다는 점에서 총 47점의 위험 점수가 부과되었지만, 직접적으로 행정적인 의사결정을 내리지 않아 전체적인 위험 수준은 2로 판단되었다. 완화 점수는 총 28점으로 집계되었는데, 해당 완화 점수는 법률 자문 등 내부 협의를 마쳤고, 연방 개인정보보호 감독기구 OPC(Officeofthe Privacy Commissioner)와 항공사 등 외부 이해관계자 협의를 거쳤으며, 책임 할당, 로그기록 및 감사추적 기능, 당사자 이의제기 및 인간의 기각 보장, 젠더 기반 분석 및 개인정보보호 영향평가 실시, 보안조치 이행 등이 이루어진 데 따른 것이다.

69) Government of Canada. Algorithmic Impact Assessment - ArriveCAN Proof of Vaccination Recognition. https://open.canada.ca/data/en/dataset/afc17416-3781-422d-a4a9-cc55e3a053c8(접근일: 2022. 8. 15).

3. 영국 인공지능 영향평가

가. 인공지능 조달지침

영국 정부는 2020년 6월 인공지능 조달지침을 발표하고 공공조달을 통하는 인공지능에서 10대 원칙을 따르도록 요구하였다.[70] 이 지침은 특히 두 가지 방식의 평가를 요구하고 있는데, 조달 절차 개시 전에는 데이터 평가를 실시하고, 조달 절차 개시 단계에서 인공지능 배치의 편익과 위험에 대하여 영향평가를 실시할 것을 요구한다.

우선 조달 기관은 데이터 평가를 통하여 △조달 절차의 개시 단계부터 데이터 거버넌스 메커니즘이 가동될 수 있도록 확보하고, △프로젝트에 관련 데이터를 사용할 수 있는지 여부를 평가하며, △시장에 출시하기 전에 데이터 내부의 결함 및 편향 가능성을 해결할 수 있어야 한다. 데이터 문제를 직접 해결할 수 없는 경우 이를 해결하기 위한 계획을 수립해야 한다. 더불어 △조달 계획 및 후속 프로젝트를 위해 공급업체와 데이터를 공유할 것인지 여부 및 방법을 정의할 것 또한 요구한다. 데이터에 대한 철저한 평가가 어려운 것으로 드러나거나 이루어지지 않은 경우, 인공지능 시스템이 의사결정의 기반으로 사용할 데이터에 대해 종합적인 점검을 실시할 것을 입찰공고 요구사항에 포함하여야 한다. 입찰 공고는 데이터를 덜 침해적으로 사용하거나 덜 민감한 데이터셋을 이용하여 동일하거나 유사한 결과를 달성하는 혁신적인 기술 접근법을 장려해야 한다.

한편, 인공지능 배치의 편익과 위험에 대한 영향평가와 관련하여 조달기관은 △제안서 평가와 의사결정에 있어 공익이 주요 요소라는 점을 조달 문서에 설명하고, 인공지능 시스템이 인간과 사회 경제에 미치는 영향 및 편익을 고려하도록 하며, 해당 조달 사업이 공익적 목표와 관련이 있고 차별금지, 동등한 대우 및 비례성의 원칙을 준수할 것을 요구한다. 더불어 △당면한 문제 해결과 관련하여 인공지능을 고려하는 배경을 조달 문서에 명확히 설명하고 대안적 솔루션에 대하여 열린 태도를 취해야 하며, △조달 절차 개시 단계에서 인공지능 영향평가를 수행하고, 중간 평가 결과가 조달에 반영되는지 확인하여야 한다. 주요 의사결정 단계에서는 평가 결과를 재차 살펴보아야 한다. 즉, 인공지

70) UK Government. Guidelines for AI procurement. <https://www.gov.uk/government/publications/guidelines-for-ai-procurement/guidelines-for-ai-procurement(접근일: 2022. 8. 15)>.

능 영향평가는 프로젝트 설계 단계에서 실시하고 이후 솔루션 설계 및 조달 절차에서 확인된 위험성의 완화를 추구해야 하며, 장래 구현될 인공지능 시스템에 대한 완전한 평가가 사실상 불가능하기 때문에 반복적으로 점검하여야 한다는 것이다.

조달지침이 제시하는 인공지능 영향평가의 항목은 6가지로서, ①인공지능 시스템에 대한 사용자 요구사항과 그 공익, ②인공지능 시스템의 인적·사회경제적 영향, 즉 인공지능이 사회적 가치 편익을 제공할 수 있도록 보장하는지, ③기존의 기술적, 절차적 환경에 미치는 결과, ④데이터 품질 및 부정확성 또는 편향성, ⑤의도하지 않은 결과가 나올 가능성, ⑥지속적인 지원 및 유지보수 요구사항 등 전체 생애주기에 대한 비용적 고려 등이다. 관련 위험성과 각각의 완화 전략이 영향평가 내에서 규정되고 합의되어야 하며, 이 전략은 '계속진행할지 중단할지' 등 주요한 의사결정을 포함해야 한다. 이러한 의사결정 또는 인공지능 시스템 설계에 상당한 변화가 있을 때마다 영향평가를 검토하여야 한다.

나. NMIP 알고리즘영향평가

보건의료 분야는 자동화된 진단에서 개인 맞춤형 의약품까지 데이터와 인공지능의 활용으로 각광받고 있지만, 다른 한편으로 민감한 개인 의료정보의 활용과 알고리즘 편향이 사생활 침해와 의료 접근의 차별 등 부정적인 영향을 미칠 수도 있다는 우려가 커져 왔다. 이에 영국에서는 보건의료 분야 인공지능 시스템의 개발 및 도입 이전에 잠재적인 해악을 식별하고 완화할 수 있는 조치를 취하기 위하여 보건의료 분야 국가의료이미지플랫폼에 특화된 알고리즘영향평가를 도입하였다.

이 영향평가도구는 에이다 러브레이스 연구소(Ada Lovelace Institute)가 국민 보건 서비스(National Health Service, 이하 'NHS') AI Lab의 지원을 받아 2021년 개발하였다. 에이다 러브레이스 연구소는 2018년 초 너필드 재단(Nuffield Foundation)이 앨런튜링 연구소, 왕립협회(Royal Society) 등과 협력하여 설립한 독립적인 연구소로서, 사람과 사회를 위한 데이터 및 인공지능의 작동을 보장하는 것을 조직의 사명으로 하고 있다. 이 연구소는 2020년 <블랙박스 검사> 보고서[71]를 발표하면서 알고리즘 감사와 알고리즘영향

71) Ada Lovelace Institute(2020). Examining the Black Box.

평가 관련 연구 및 실무 현황을 검토한 바 있다. 보고서에 따르면 알고리즘영향평가는 1) 시스템이 사용되기 전에 알고리즘 시스템의 사회적 영향 가능성을 평가하는 ‘알고리즘 위험평가’와 2) 알고리즘 시스템이 사용자나 인구집단에 미칠 수 있는 사회적 영향을 평가하는 ‘알고리즘영향평가’의 두 가지 접근법으로 나누어 살펴 볼 수 있다.

이러한 배경에서 연구소는 보건의료 알고리즘영향평가 프로젝트에 착수하여 어떤 인공지능 시스템이 NHS AI Lab의 국가의료이미지플랫폼(National Medical Imaging Platform, 이하 ‘NMIP’)의 데이터를 활용하고자 할 때 그 영향을 평가하는 도구를 개발하였다. NMIP는 고화질 흉부 X-레이, MRI, 피부, 안과 등 대규모 데이터셋으로서, 연구자나 민간기업이 의료 인공지능 제품을 실험, 학습, 검증하는데 제공하기 위한 목적으로 만들어진 것이다. NMIP의 데이터셋에 접근하고자 하는 연구자는 알고리즘영향평가(Algorithmic Impact Assessments, 이하 ‘NMIP 알고리즘영향평가’) 도구를 통해 사전에 영향평가를 진행해야 하며, 그 결과에 따라 NMIP 데이터를 제공할 것인지 여부가 결정된다. 이를 위해 알고리즘영향평가도구인 <NMIP 알고리즘영향평가 템플릿>[72]이 고안되었으며, 연구자가 NMIP 알고리즘영향평가의 모든 절차를 따라할 수 있도록 <NMIP 알고리즘영향평가 사용자 가이드>[73]가 제공된다.

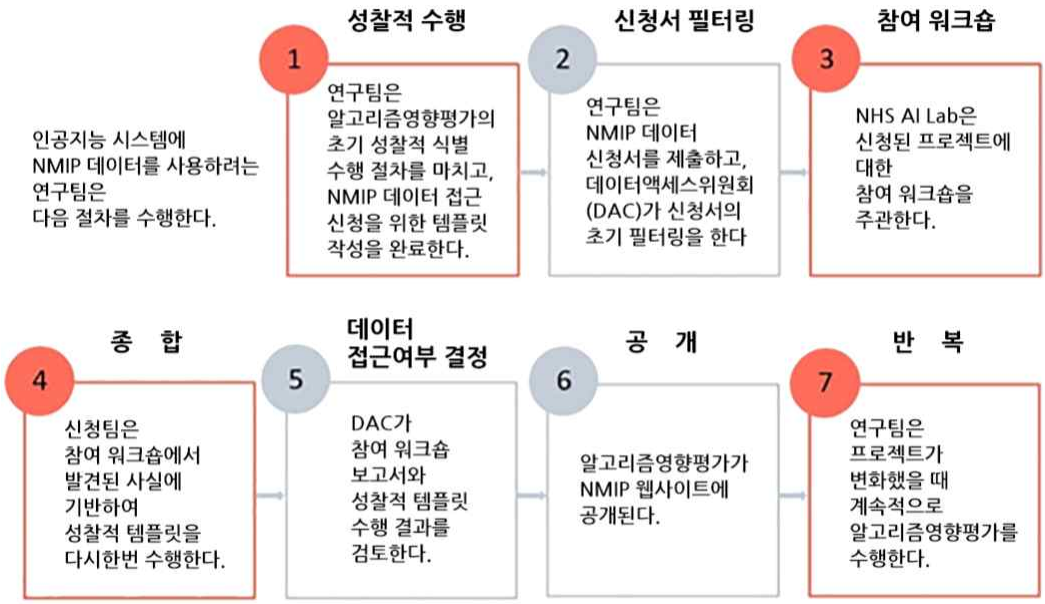
연구소는 알고리즘영향평가를 “알고리즘 시스템을 사용하기 전에 해당 시스템의 잠재적인 사회적 영향을 평가하기 위한 도구”로 규정하는데, 여기서 ‘인권’에 대한 영향뿐 아니라 보다 폭넓은 사회적 영향에 대한 평가를 목표로 하고 있음을 알 수 있으며, 이는 템플릿의 영향평가를 위한 질의를 통해서도 확인할 수 있다.

<https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf(접근일: 2022. 8. 15)>.

72) Ada Lovelace Institute(2020). 번역은 부록 III 참조.

73) Ada Lovelace Institute(2020). Algorithmic impact assessment: user guide. <https://www.adalovelaceinstitute.org/resource/aia-user-guide/(접근일: 2022. 8. 15)>.

[그림 7] 영국 NMIP 알고리즘영향평가 흐름



1) 평가의 절차

NMIP 알고리즘영향평가는 아래 그림과 같이 7개의 절차로 이루어진다. NMIP 데이터에 접근하고자 하는 연구팀은 이러한 절차를 차례대로 수행해야 한다.

○ 1단계: 성찰적 수행(AIA reflexive exercise)

NMIP 알고리즘영향평가 절차는 NMIP에 대한 접근 신청서를 제출하기 전에 시작된다. 프로젝트에 대한 정보, 신청자의 인공지능 시스템으로 인해 발생할 수 있는 피해와 영향, 영향을 받은 이해관계자 등을 식별하기 위한 것으로, 템플릿의 도움을 받아 신청자팀 스스로 작성하도록 하고 있다. 템플릿은 프로젝트에 대한 설명과 함께, 시스템이 미치는 영향을 평가하기 위한 일련의 질의와 답변으로 구성된다.

영국 NMIP 알고리즘영향평가 템플릿 질의 예시[74)]

2. 통상적인 윤리적 고려 사항

이 섹션에서는 의료, AI 및 알고리즘 문헌의 맥락에서 일반적인, 특정 윤리적 고려 사항을 안내합니다.
이 섹션 작성에 도움이 필요한 경우, NHSX의 AI 보고서 섹션 3을 참조하여 의료 분야에서 AI를 사용하기 위한 윤리적 고려 사항뿐만 아니라, 보다 구체적으로는 데이터/디지털 건강 및 알고리즘 고려 사항을 참조할 수 있습니다. 이는 직간접적인 이해 관계자에 대한 당신의 모델의 가능한 영향을 분석하는 데 도움이 될 것입니다. 예를 들어, 프롬프트 2b에서, 당신은 사생활과 데이터 공유, 그리고 환자 안전에 대한 가능한 영향에 대해 논의할 수 있습니다.

2.a 이 프로젝트가 특정 커뮤니티에 대한 불평등 또는 불법적인 차별의 생성 또는 악화로 이어질 수 있습니까? 예를 들어, 치료에 대한 차별적 접근을 악화시키면서? 편향 및 공정성을 평가하거나 모니터링하기 위한 현재 계획에서 간과할 수 있는 것은 무엇입니까?

성찰적 수행	참여 워크숍	종합

2.b 귀하의 프로젝트는 동의와 자율성을 어떻게 고려합니까? 감시 증가와 관련된 위험이 있습니까? 예를 들어, 시스템의 의도된 수혜자에게 시스템 사용에 대해 어떻게 알립니까? 이 시스템은 감시가 증가하는 것으로 해석될 수 있습니까?

성찰적 수행	참여 워크숍	종합

○ 2단계: 신청서 필터링

신청자가 완료된 템플릿을 NMIP의 데이터액세스위원회(Data Access Committee, DAC)에 제출하며, DAC는 이를 심사한다. 심사기준을 충족하는 신청자는 3단계로 진행한다.

○ 3단계: 참여 워크숍(participatory workshop)

NHS AI Lab에서 신청된 프로젝트에 대한 참여 워크숍을 주관한다. 의사, 환자 등 시스

74) Ada Lovelace Institute(2020). 전체 질의 번역은 부록 III 참조.

템의 영향을 받은 다양한 이해관계자들이 참여 워크숍에 참여하며, 해당 프로젝트의 잠재적인 영향에 대해 논의한다. NHS가 지정하는 보고관도 워크숍에 참석하여 DAC가 최종 결정에 참고할 수 있도록 보고서를 작성한다.

참여 워크숍은 NMIP 알고리즘영향평가 절차에 이해관계자의 참여를 보장함으로써 책무성과 투명성을 높일 수 있도록 한다. 워크숍은 연령, 성별, 지역, 민족적 배경, 사회경제적 배경, 건강 상태 또는 치료 접근성에 걸쳐 알고리즘의 영향을 받을 수 있는 인구의 다양성을 반영하는 패널 8-12명으로 구성된다. 또한 기술 및 사회 전문가가 비판자(critical friend)로 참여하도록 하고 있다. 신청자 역시 워크숍에 참여하여 프로젝트에 대해 설명하고 패널의 질문에 답변하며 토론을 경청한다. 사용자 가이드는 부록에서 참여 워크숍에 대한 구체적인 진행 사례를 제시하고 있다.

○ 4단계: 종합(Synthesis)

참여 워크숍이 끝나면 신청자는 워크숍에서 깨달은 내용을 기반으로 템플릿을 업데이트한다. 이를 위해 1단계 성찰적 수행을 함께 수행한 팀 구성원을 다시 소집할 수 있다.

○ 5단계: 데이터 접근여부 결정

업데이트된 NMIP 알고리즘영향평가 템플릿과 참여 워크숍에 대한 보고자의 요약 보고서가 DAC에 제출된다. DAC는 NMIP 알고리즘영향평가의 내용을 평가하고 기타 자료를 검토하여, NMIP 데이터에 대한 접근을 허용할지 여부를 결정한다.

이 제안에서는 DAC를 사회과학, 생의학, 컴퓨터 과학, 법학을 전문으로 하는 학계 전문가, 환자단체 대표 등 최소한 11명 이상으로 구성할 것을 권고하고 있다.

○ 6단계: 공개

NMIP 알고리즘영향평가 결과물은 NHS 중앙 저장소에 공개되는데, 이 영향평가에 대해 문의할 수 있도록 신청자의 연락처도 함께 공개된다. 심사를 통과한 성공적인 영향평가만 공개되지만, 그렇지 않은 신청자 역시 결과물을 공개하여, 자신이 배운 내용을 공유할 수 있다.

○ 7단계: 반복

NMIP 알고리즘영향평가는 주기적으로 반복될 수 있는데, 여기에는 새로운 팀 구성원이 이 절차에 참여하고 추가적인 성찰을 위한 2년 마다의 정기적인 평가가 포함된다. DAC가 재량으로 재평가를 제안할 수도 있다. 인공지능 시스템이 제품의 기능, 범위 및 애플리케이션의 변경, 사용자 기반의 변화와 같이 중대한 변경이 이루어질 경우 영향평가를 다시 수행해야 한다.

2) 평가의 기준

성찰적 수행에 사용되는 템플릿은 4개의 섹션으로 구성된다.

첫째는 프로젝트에 대한 상위 수준의 정보를 제공한다. 프로젝트의 목적과 의도된 용도, 프로젝트를 수행하는 조직, 시스템 및 모델의 입력 및 결과물에 대한 세부정보(예를 들어, 어떤 종류의 데이터 소스를 사용하는지), 그리고 시스템의 영향을 받는 이해관계자에 대한 정보가 포함된다. 이해관계자의 경우 의도된 사용자는 누구인지, 서비스 대상은 무엇인지 등을 임상의, 간호사, 병원행정직원, 영향을 받는 인구집단 등을 구체적으로 제시하도록 하고 있다.

둘째는 일반적인 윤리적 고려사항에 대한 질문이다. 특정 커뮤니티에 대한 차별 가능성, 동의와 자율성, 감시와 관련된 위험, 환경에 미치는 영향, 환자와 치료 전문가의 관계에 미치는 영향, 영향을 받는 당사자가 결과에 이의를 제기할 수 있는 방법, 의도하지 않은 오류 혹은 의도적인 오용의 가능성 등을 평가하도록 한다.

셋째는 영향 식별 및 시나리오 섹션으로 최상의 시나리오는 무엇인지, 성공적 운영을 위한 사회환경적 요구조건, 그리고 과제나 장애물은 무엇인지, 최악의 시나리오의 경우 시스템이 의도된 대로 작동할 때 및 그렇지 않을 때 발생할 수 있는 상황은 무엇인지 등을 분석하도록 한다.

넷째, 잠재적 피해 분석 섹션은 모든 시나리오에서 발생하는 다양한 이해관계자의 잠재적 피해가 무엇인지, 그리고 피해를 최소화하기 위한 완화조치는 무엇인지 분석하도록 한다. 피해를 분석할 때에는 피해의 중요도, 긴급성, 완화조치의 어려움, 탐지가능성을 고려하도록 한다.

3) 다른 규제 메커니즘과의 관계

보건의료 분야에서는 공공성의 보장 및 인권 보호를 위한 여러 규제 메커니즘이 존재한다. 연구소는 NMIP 알고리즘영향평가가 기존의 규제 메커니즘과 대체되거나 중복적인 것이 아니라 상호 보완적이라고 설명한다. 즉, 규제 대상이 되는 기관들은 NMIP 알고리즘영향평가와 무관하게 개인정보보호 영향평가, 의약품 기기 위험 분류 등의 규제를 준수해야 한다. 그러나 NMIP 알고리즘영향평가는 의료 인공지능 개발에 다양한 이해관계자들이 참여할 수 있도록 돕고, 다른 규제 메커니즘에서 다루어지지 않는 위험 및 영향(예를 들어 사회적인 영향)을 다룰 수 있으며, NMIP 알고리즘영향평가의 결과물을 문서화하고 투명하게 공개하는 표준화된 방법을 제공한다. 아래 그림은 영국 보건의료 분야의 다양한 규제 메커니즘 내에서 NMIP 알고리즘영향평가가 어디에 위치하는지 보여준다.

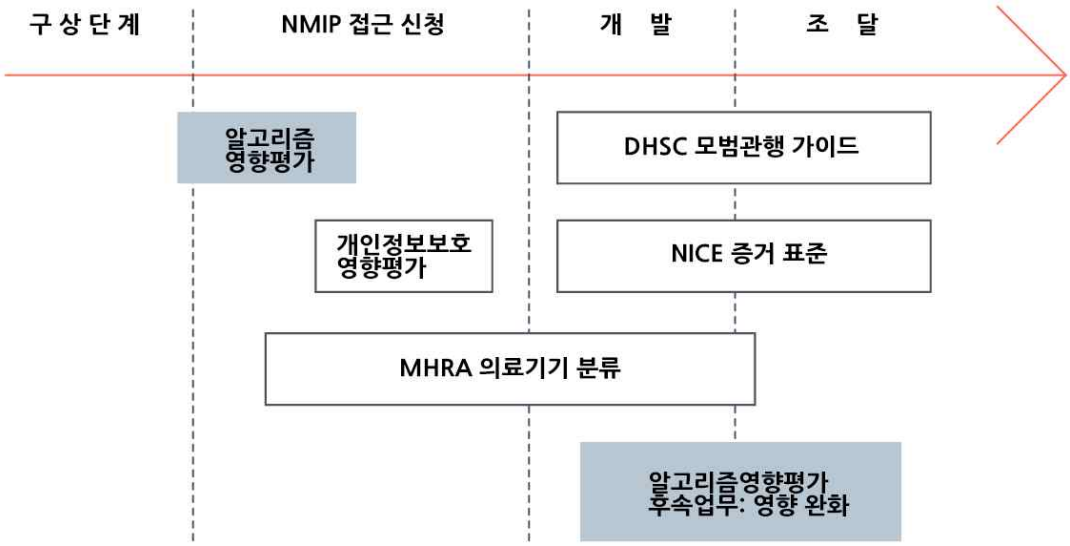
4) 운영상 쟁점

에이다 러브레이스 연구소는 프로젝트의 최종 보고서를 발간하였는데, 이 보고서에서는 보건의료 분야의 알고리즘영향평가 사례를 다른 맥락에 적용할 경우 고려해야 할 운영상의 쟁점들을 제시하고 있다.

(가) 영향평가 생태계

환경영향평가 등 다른 영향평가와 달리, 알고리즘영향평가의 경우 아직 일관된 방법론이 없고 법적 근거도 부족하며, 이를 수행할 수 있는 평가자 시장도 존재하지 않는다. 더구나 인공지능 시스템은 의료, 금융, 사법, 공공 서비스 등 다양한 맥락에서 사용될 수 있으므로, 해당 맥락에 맞게 질문의 범위를 조정할 필요가 있다. NMIP 사례에서는 DAC를 평가자의 중심에 두고, 데이터와 인공지능 시스템의 윤리적 사용에 대한 기존 NHS 지침을 차용하는 방식으로 이 문제를 해결하였다.

[그림 8] 영국 NMIP 알고리즘영향평가와 다른 규제 메커니즘의 관계



(나) 평가 기초 작업

NMIP 사례에서, 알고리즘영향평가는 보건의료 분야에서 이미 사용되고 있는 품질 및 기술 보증 또는 위험 관리를 위한 다른 메커니즘을 대체하지 않는다. 알고리즘영향평가는 인공지능 시스템 관리를 위한 완전한 해결책으로 이해되어서는 안되며, 영향평가를 시행하는 구체적인 맥락에서 '사전' 혹은 '사후'의 작업을 고려해야 한다. 데이터세트에 대한 좋은 문서를 확보하는 것도 중요하다. NMIP 사례에서는 모두 동일한 데이터를 사용하지만, 다른 맥락에서는 데이터가 어떤 동의 수준에서 수집되었는지, 데이터의 출처와 형식, 편향 테스트를 했는지 여부 등 데이터셋을 둘러싼 문서화가 필요하다.

(다) 평가 수행 주체

기존의 성공적인 영향평가 모델에서는 독립적인 평가자의 중요성을 강조한다. 설명 책임(책무성)이 중요할 경우 평가의 독립성이, 성찰성에 중점을 둘 경우에는 시스템 개발자의 학습과 개선을 위한 메커니즘으로 자체 평가를 우선시할 수 있다. NMIP의 사례에서는 개발자를 위한 성찰적인 절차를 허용하고, DAC가 독립 평가자로서 이를 검토하는 절차를 둠으로써 두 가지 관심사를 모두 포착하려 하고 있다.

(라) 이해관계자 참여

알고리즘영향평가에 대한 여러 문헌은 다학제 이해관계자, 영향을 받는 지역사회, 일반 대중의 의견 수렴과 참여를 제안하는데, 이는 개발자와 시스템의 영향을 받는 사람들 사이의 책임있는 관계를 위해 중요하다. 그러나 누구를 참여시킬 것인지와 함께, 참여의 유형도 정보제공에서부터, 협의, 합의 구축을 위한 협업까지 다양할 수 있다. 어떠한 참여 실행은 명목적이거나 형식적일 수 있으며, 이 문제를 해결하기 위해서는 참여자가 자유롭게 심의하고 비판적 피드백을 제공할 수 있도록 해야하며, 자신이 제기한 우려를 개발자가 다룰 것이라 확신할 수 있어야 한다. 또한 규모나 특성에 맞게 유연하고 확장가능해야 한다.

(마) 평가 결과 공개

알고리즘영향평가의 결과물을 다른 사람들이 검토하고 평가할 수 있도록 공개할 필요가 있다. 그러나 이를 위한 표준화된 형식이나 템플릿은 존재하지 않는다. 영향평가 결과물을 공개할 때 다음과 같은 사항을 고려해야 한다. 첫째, 게시되는 내용. NMIP 사례에서는 개발자가 사용하는 템플릿 결과물을 공개한다. 둘째, 게시되는 장소. NMIP 사례에서는 NMIP 데이터를 사용하는 신청자에 의해 수행되므로, NMIP가 중앙 허브 역할을 하고 있으며, 공공 부분에서는 중앙 집중화된 등록소가 그러한 역할을 할 수 있다. 셋째, 공개로 인한 위험성. 이러한 공개가 지적재산권이나 상업적 민감성 문제를 제기할 수 있고, 성찰성보다는 홍보에 대한 동기를 유발할 수 있다는 우려도 제기되지만, 영향평가의 더 큰 목표와 균형을 이룰 필요가 있다.

(바) 영향평가의 효과성 평가 및 후속조치

특정한 맥락에서 알고리즘영향평가의 효과를 평가하기 위한 표준이 아직 없으며, 서로 다른 알고리즘영향평가의 효과성을 평가하기도 어렵다. 시스템의 부정적 영향을 문서화하고서도 개발팀이 시스템의 개발 및 보급을 진행하거나, 인공지능 영향평가를 제대로 수행하지 않고 단지 지속적인 개발을 위한 명분으로 영향평가 결과를 활용한다면, 이는 알고리즘영향평가가 제대로 역할하지 못한 것이다.

또한, 알고리즘영향평가는 어떠한 부정적 영향을 포착했을 때 무엇을 할 것인지에 대

한 결정을 요구한다. 즉 영향이 크지 않아 조치를 취하지 않을 것인지, 완화 조치를 적용하여 시스템의 일부를 수정할 것인지, 개발 및 사용을 중단할 것인지 등을 결정해야 하는데, 누가 이러한 결정권을 가질 것인지 정해야 한다.

(사) 영향평가 자원 및 지원

알고리즘영향평가는 제안 단계이며 실제 효과가 있는지 알 수 없다. 따라서 알고리즘영향평가를 수행하려는 사람들은 이 절차를 시험하고 평가하며, 절차를 반복할 것을 고려해야 한다. 이를 위해 전문성과 같은 자원, 자금 지원, 실험에 대한 평가가 필요하다.

4. 미국 알고리즘영향평가

2022년 10월 미국 백악관 과학기술정책국은 <인공지능 권리장전 청사진>을 발표하였다.[75] 이 청사진은 기업과 정부 기관들의 인공지능이 준수하여야 할 5가지 원칙으로 ① 안전하고 효과적인 시스템, ②알고리즘 차별로부터 보호, ③개인정보 보호, ④통지 및 설명, ⑤인간 대안・검토 및 대체를 제시하면서 각 분야 위험을 식별하고 완화하기 위하여 영향평가의 수행과 공개를 강조하였다.

첫째, 안전하지 않거나 효과적이지 못한 시스템으로부터 국민을 보호하기 위해서는 시스템이 안전하고 효과적이라는 사실을 확인하는 독립적인 평가와 보고가 수행되어야 하며, 여기에는 잠재적 위해를 완화하기 위해 취한 조치에 대한 보고가 포함되어야 하고, 그 결과는 가능할 한 공개되어야 한다. 둘째, 알고리즘 차별로부터 국민을 보호하기 위해서, 시스템 설계, 대표 데이터 사용 및 인구집단 특성별 대리변수 보호에 대한 공평성 평가를 사전에 수행하고, 설계 및 개발에 대한 장애인에 대한 접근성을 보장하며, 배치 전 및 지속적으로 편향성 테스트 및 완화조치를 수행하고, 조직적 감독을 명확하게 수행하여야 한다. 더불어 알고리즘영향평가 형식의 독립적인 점검과 쉬운 용어로 이루어진 보고가 이루어져야 하며, 여기에는 편향성 테스트 결과 및 완화 조치에 대한 정보가 포함되어야 하고, 이러한 보호 조치 여부를 확인하기 위해 보고서가 가능한 한 공개되어야

75) The Office of Science and Technology Policy(2022). 번역은 부록 II 참조.

한다. 셋째, 개인정보 보호를 위해서, 감시 기술이 사생활과 시민권에 미치는 잠재적 위해와 이를 제한하는 범위를 최소한 배치 전에 평가하는 등 감시 기술에 대한 감독을 강화하여야 한다. 국민은 자신의 개인정보 결정권이 존중되었음을 확인하고 자신의 권리, 기회 또는 접근에 미치는 감시 기술의 잠재적 영향에 대해 평가하는 보고서에 접근할 수 있어야 한다. 넷째, 통지 및 설명을 보장하기 위해서도, 자동화된 시스템에 대한 요약 정보가 쉬운 용어로 포함된 보고서와 더불어, 이들 통지 및 설명의 명확성과 품질에 대한 평가 내용이 가능한 한 공개되어야 한다. 다섯째, 인간 대안·검토 및 대체가 가능하기 위해서는 인간 거버넌스 절차에 대한 설명과 적시성, 접근성, 결과물 및 효과성에 대한 평가 보고서가 가능한 한 공개되어야 한다.

한편, 2022년 2월 3일 미국 하원과 상원에 <알고리즘 책무성법(안)>[76]이 함께 발의되었다. 이 법안은 연방거래위원회(FTC)가 소관하는 일정 규모 이상의 기업들을 대상으로 자동화된 의사결정 시스템 또는 증강된 중요 의사결정 프로세스 및 이들이 소비자에게 미치는 영향에 대하여 지속적으로 연구·점검하는 '영향평가'(§2(12))를 실시하도록 의무화하고 FTC가 감독하도록 하였다. 이 법안을 앞서 유럽연합 인공지능(안)과 비교하여 보았을 때 용어 면에서는 '인공지능'이라는 용어 대신 '자동화된 의사결정 시스템 또는 증강된 중요 의사결정 프로세스'라는 용어를 취하였고, '고위험'이라는 용어보다 '중요 의사결정'이라는 용어를 사용하였다. 이 법안은 대상 면에서 법집행, 출입국, 사법 분야 등 공공부문을 부분적으로 포함하고 있는 유럽연합 인공지능법(안)과 차이가 있으며, 2017년 뉴욕시가 미국에서 처음으로 입법한 일명 「뉴욕시 알고리즘 책무성법」[77]이 공공기관의 자동화된 의사결정 시스템에 적용되는 것과도 차이가 있다.

알고리즘 책무성법(안)은 자동화된 의사결정 시스템에 대한 영향평가의 경우, 평균 연수입이 5천만 달러를 초과하거나, 3년간 지분 가치가 2억5천만 달러를 초과하거나, 소비자, 가구 또는 소비자 장치에 대한 식별 정보를 1백만 건 이상 처리하는 개인, 파트너 및 회사가 대상이며(§2(7)(A)(i)), 배치 전에 영향평가를 실시하도록 하였다(§3(b)(1)(A)(i)). 증강된 중요 의사결정 프로세스에 대한 영향평가의 경우, 평균 연수입이 5백만 달러를

76) Algorithmic Accountability Act of 2022, H.R.6580(하원발의안) 및 S.3572(상원발의안).

77) A Local Law in relation to automated decision systems used by agencies, Int. No. 1696-A.

초과하거나 3년간 지분 가치가 2천5백만 달러를 초과하는 개인, 파트너 및 회사로 그 대상이 확대되며(§2(7)(A)), 배치 전후에 실시되어야 한다(§3(b)(1)(A)(ii)). 여기서 '중요 의사결정'이란 다음과 같은 서비스에 대한 접근 또는 비용·조건·가용성과 관련하여 소비자의 삶에 법적으로나 중요한 또는 유사하게 중대한 영향을 미치는 의사결정 또는 판단을 의미한다(§2(8)).

미국 알고리즘 책무성법(안) 중요한 의사결정(§2(8))

(A) 평가, 인정, 인증을 포함한 교육 및 직업훈련
(B) 고용, 근로자 관리, 자영업
(C) 전기, 난방, 수도, 인터넷·통신 접근, 교통과 같은 필수 설비
(D) 입양 서비스, 생식 서비스를 포함한 가족 계획
(E) 모기지 회사, 모기지 브로커, 채권자가 제공하는 금융 서비스를 포함한 모든 금융 서비스
(F) 정신건강의학과, 치과, 안과를 포함한 모든 보건의료
(G) 주택 임대, 단기 주택 임대, 숙박 서비스를 포함한 모든 주택 및 숙박
(H) 사적 중재 또는 조정을 포함한 법률 서비스
(I) 위원회가 규칙 제정을 통해 소비자의 삶에 비교적 법적, 물질적 또는 유사하게 중대한 영향을 미친다고 판단한 서비스, 프로그램 또는 기회 결정

대상 기업은 수행된 영향평가 관련 문서를 시스템 또는 프로세스 배치 후 3년 이상 보관하여야 하며, 영향평가에 대한 요약보고서를 매년 FTC에 제출하는 한편, 신규 시스템 또는 프로세스의 경우 초기 요약보고서를 그 배치 전에 제출하여야 한다. 대상 기업은 영향평가 수행시 관련 내부 이해관계자(직원, 윤리 팀 및 담당 기술팀 등) 및 독립적인 외부 이해관계자(영향을 받는 집단의 옹호자나 대표, 시민 사회 및 인권단체, 기술 전문가 등)와 필요에 따라 수시로 의미 있는 협의(참여 설계, 독립 감사 또는 피드백 요청 및 통합)를 하여야 한다. 소비자의 삶에 법적 또는 유사하게 중대한 영향을 미치는 물질적·부정적 영향이 나타나는 경우 프로세스에 의해 발생하는 모든 영향을 시기적절한 방식으로 제거하거나 완화하기 위해 노력하여야 한다(§3(b)(1)).

영향평가 요구사항은 법률로 구체적으로 규정되어 있으며(§4(a)), ①프로세스를 새로 도입하는 경우 기존 프로세스에 대한 검토 ②이해관계자 협의 ③개인정보보호 테스트 및

검토 ④현재 및 과거 성능에 대한 지속적인 테스트 및 검토 ⑤직원에 대한 지속적인 교육훈련 ⑥시스템 또는 프로세스의 특정한 사용 및 적용에 대한 보호막이나 한정의 필요성과 개발 가능성 ⑦개발, 테스트, 유지 관리, 갱신하는 데 사용되는 데이터 및 기타 입력 정보에 대한 최신 문서의 유지 관리 및 보관 ⑧소비자의 권리 ⑨소비자에게 미치는 중대한 부정적 영향 가능성의 식별 및 적용가능한 완화 전략 ⑩개발 및 배치 절차에 대하여 진행 중인 문서화 ⑪개선이 필요한 기능, 도구, 표준, 데이터셋, 보안 프로토콜, 이해관계자 참여 및 기타 자원 ⑫미준수 영향평가 요구사항 및 미준수 근거 ⑬위원회가 적절하다고 판단한 기타 진행 중인 연구 또는 검토로 구성된다.

이처럼 알고리즘 책무성법(안)은 영향평가 요구사항을 통해 시스템의 기준 준수를 유도한다는 점에서 유럽연합 인공지능법(안)에 비하여 실용적인 접근을 취하고 있다고 평가된다. 다만 영향평가의 구체적인 정보와 집행 범위가 모두 관할기관인 FTC에 위임되어 있어 이 법의 규범력이 명확하지 않다는 점이 문제로 지적되고 있다.[78)]

[표 4] 미국 2022년 알고리즘 책무성법(안) 영향평가 요구사항

	요구사항	세부항목
1	프로세스를 새로 도입하는 경우 기존 프로세스에 대한 검토	(A) 신규 프로세스로 개선 또는 대체되는 기준 프로세스 설명 (B) 기존 프로세스의 알려진 위해, 단점, 오류 사례 및 소비자에게 미치는 중요한 부정적 영향 (C) 프로세스에서 의도하는 편익 및 필요성 (D) 시스템 또는 프로세스가 의도하는 목적
2	이해관계자 협의	(A) 협의 이해관계자 연락처 (B) 협의 일자 (C) 협의 조건 및 절차에 대한 정보 (i) 이해관계자와 대상 기업 간 법적 또는 재정적 합의의 존재 유무 및 성격 (ii) 이해관계자와 상호교류한 모든 데이터, 시스템, 설계, 시나리오 및 기타 문서와 자료 (iii) 시스템 또는 프로세스의 개발 또는 배치를 변경하는 데 사용된 이해관계자의 모든 권장사항, 사용되지 않은 권장사항 및 미사용 근거

78) Mökander, J., Juneja, P., Watson, D.S. et al(2022). The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: what can they learn from each other?. Minds & Machines 32, pp.751-758.

3	개인정보보호 테스트 및 검토	(A) 데이터 최소화 관행 및 관련 식별 정보와 결과적으로 내려진 중요 결정이 저장되는 기간 (B) 연합 학습(federated learning), 차등 개인정보보호, 안전한 다자간 계산, 비식별화, 위험 수준에 기반한 보안 데이터 영역 등 개인정보보호 강화 기술의 사용 등 정보 보안 조치 (C) 소비자와 식별 정보의 개인정보보호, 안전 및 보안에 대하여 현재 및 잠재적 미래, 또는 외부적으로 미칠 수 있는 긍정적 및 부정적 영향
4	현재 및 과거 성능에 대한 지속적인 테스트 및 검토	(A) 대상 기업이 성공적인 수행과 기법으로 간주하는 사항과, 성능을 평가하기 위해 사용한 기술 및 사업 측정기준(metrics) (B) 테스트 조건에서 해당 시스템 또는 프로세스의 성능에 대한 검토, 또는 이러한 성능 테스트가 수행되지 않은 이유 설명 (C) 배치 조건에서 해당 시스템 또는 프로세스의 성능에 대한 검토, 또는 배치 조건에서 성능이 검토되지 않은 이유 설명 (D) 배치 조건에서 해당 시스템 또는 프로세스의 성능을 테스트 조건과 비교하거나, 이러한 비교가 불가능한 이유 설명 (E) 소비자의 인종, 피부색, 성별, 성정체성, 연령, 장애, 종교, 가족 상태, 사회경제적 상태, 병역 퇴역 상태, 또는 위원회가 적절하다고 판단하는 기타 특성(해당 특성의 조합을 포함)과 관련된 모든 차별적인 수행에 대한 검토. 이러한 검토를 위한 방법론에 대한 설명, 데이터에서 해당 특성을 식별하는 데 사용한 방법(우편번호를 비롯한 대리적인 데이터의 사용 등) (F) 테스트 및 검토에 하위 집단이 사용된 경우, 사용된 하위 집단과 해당 하위집단이 테스트 및 검토와 관련이 있다고 판단된 방법 및 이유
5	피고용인, 계약직 및 기타 직원에 대하여 시스템 또는 프로세스가 소비자에게 미치는 중대한 부정적 영향 및 개선 등에 대하여 지속적인 교육훈련	
6	시스템 또는 프로세스의 특정한 사용 및 적용에 대한 보호막이나 한정의 필요성과 개발 가능성	
7	개발, 테스트, 유지 관리, 갱신하는 데 사용되는 데이터 및 기타 입력 정보에 대한 최신 문서의 유지 관리 및 보관	(A) 해당 데이터 및 기타 입력 정보를 취득한 시기 및 방법, 라이선스가 부여되었는지 여부 (i) 파일 유형, 파일 생성・수정일, 데이터 필드 설명 등 메타데이터 (ii) 대상 기업이 데이터 및 기타 입력 정보를 수집, 추론, 획득한 방법론. 해당 데이터 및 기타 입력 정보에 라벨링, 분류, 정렬, 군집화를 적용한 방법론 (iii) 소비자가 자신에 관한 데이터 및 기타 입력 정보의 포함 및 추가 사용에 대하여 설명에 입각한 동의를 제공했는지 여부 및 그 방법 (B) 해당 데이터 및 기타 입력 정보가 사용된 이유 및 다른 대안이 탐색되었는지 여부

		(C) 데이터 및 기타 입력 정보에 대한 정보 (i) 데이터셋의 대표성 및 프로세스가 배치되는 인구집단 분포에 대한 가설 등 해당 요소가 측정된 방법 (ii) 데이터 품질, 그 품질의 검토 방법, 데이터를 정규화, 수정 또는 정제하기 위해 취한 조치
8	소비자의 권리	(A) 소비자에게 다음 사항을 제공하는 정도 (i) 해당 시스템 또는 프로세스가 사용될 것이라는 명확한 통지 (ii) 이러한 사용에서 제외(opt-out)될 수 있는 방법 (B) 해당 시스템 또는 프로세스의 투명성과 설명 가능성 평가. 소비자가 결정에 대해 이의제기・정정・재심을 청구하거나 해당 시스템 또는 프로세스에서 제외될 수 있는 정도 평가 (i) 그 변경 시 시스템 또는 프로세스가 다른 결정을 내리도록 하는 기여 요인 설명 등 시스템 또는 프로세스에 대해 소비자 또는 소비자의 대표자 및 대리인이 이용할 수 있는 정보 (ii) 해당 시스템 또는 프로세스와 관련하여 소비자가 대상 기업에 제출한 불만, 분쟁, 정정, 재심, 제외 요청에 대한 문서 (iii) 소비자의 우려나 피해를 해결하기 위해 대상 기업이 취한 모든 시정 조치의 과정 및 결과 (C) 제3자 결정 수취인이 해당 시스템 또는 프로세스의 결과에 대한 사본을 받거나 이에 접근할 수 있는 범위
9	소비자에게 미치는 중대한 부정적 영향 가능성의 식별 및 적용가능한 완화 전략	(A) 영향을 식별하고 측정하기 위해 취한 조치 등 소비자에게 미치는 시스템 또는 프로세스의 중대한 부정적 영향 가능성의 식별과 측정 (B) 시장에서 시스템이나 프로세스를 철수하거나 개발을 종료하는 등의 조치를 포함하여 식별된 중대한 부정적 영향 가능성을 제거하거나 합리적으로 완화하기 위해 취한 조치 (C) 식별된 중대한 부정적 영향이 완화되지 않은 상태로 남아 있는 경우 조치를 취하지 않은 이유 (D) 소비자에 미치는 중대한 부정적 영향 가능성을 식별, 측정, 완화 또는 제거하는 데 사용되는 표준 프로토콜 및 관행, 직원 교육 방법
10	개발 및 배치 절차에 대한 문서화 상황	(A) 테스트일, 배치일, 라이선스 부여일, 기타 중요한 일정 (B) 관련 부서, 사업단위 및 이와 유사한 내부 이해관계자의 연락처
11	개선이 필요한 기능, 도구, 표준, 데이터셋, 보안프로토콜, 이해관계자	(A) 정확성, 견고성 및 신뢰성을 포함하는 성능 (B) 편향성[방지]과 차별금지를 포함하는 공정성 (C) 투명성, 설명가능성, 이의제기가능성 및 소구 기회 (D) 개인정보보호 및 보안 (E) 개인 및 공공의 안전 (F) 효율성과 적시성

	참여 및 기타 자원	(G) 비용 (H) 위원회가 적절하다고 판단하는 기타 영역
12	미준수 영향평가 요구사항 및 미준수 근거	(A) 다른 사람, 제휴자, 기업이 개발한 자동화된 의사결정 시스템의 특정 정보가 부재함 (B) 대상, 고객, 라이선스 사용권자, 파트너 및 기타 개인, 제휴자 또는 기업이 증강된 중요 의사결정 프로세스에 자동화된 의사결정 시스템을 배치하는 방법에 대한 특정 정보가 부재함 (C) 해당 데이터가 수집, 추론 또는 저장하기에는 너무 민감하기 때문에 차별적 수행을 평가하는 데 필요한 인구집단 및 기타 데이터가 부족함 (D) 기술 혁신을 포함하여 해당 요구사항을 수행하는 데 필요한 특정 기능이 부족함
13	위원회가 적절하다고 판단한 기타 진행 중인 연구 또는 검토	

제2절 인공지능 인권영향평가 사례

1. 유엔 인권규범과 인공지능 인권실사

유엔 인권기구들은 인공지능 등 신기술이 인권에 미치는 부정적인 영향을 식별·방지·완화하기 위하여 인권실사의 시행을 권고하여 왔으며, 인권영향평가는 인권실사의 유용한 도구로서 인권에 미치는 부정적 영향을 식별하고 대처하는 데 도움이 된다고 여러 차례 강조하였다.

유엔 의사표현의 자유 특별보고관은 2018년 8월 보고서[79]에서 인공지능이 인권에 주는 위험으로부터 인권을 보장해야 하는 국가의 의무와 기업의 책임을 강조하였다. 특별보고관은 국가와 기업 모두에 대하여 인공지능 시스템의 조달, 개발 또는 사용에 앞서 인권영향평가를 수행하여야 하고, 이 평가는 자체 평가와 외부 검토를 모두 포함해야 한다고 권고하였다. 또한 인공지능 시스템에 대한 독립적인 외부 감사는 사전적인 인권영향평가를 보완하여 인공지능 시스템의 투명성과 책무성을 확보할 수 있다. 따라서 국가가 인공지능 시스템을 조달하거나 도입하기에 앞서 인권영향평가 또는 공공기관 알고리즘영향평가를 수행하고 도입 후에는 외부 독립전문가의 정기적인 감사를 받을 것을 권고하였다. 기업은 인공지능 제품과 서비스 출시 전에 인권영향평가를 실시하는 한편 소수자 이용자 대표 및 인권시민사회가 참가하는 공개 협의를 갖고 그 결과를 공개하여야 한다.

유엔 인권최고대표는 2020년 <최종 사용에서 인권 위험 식별 및 평가(B-테크 기초 자료)>[80]를 발표하였다. 이 문서는 기술 기업들에 대하여 유엔 기업과 인권 이행지침의 기본적인 요구를 이행할 것을 요구하였다.

나아가 유엔 인권최고대표는 2021년 <디지털 시대 프라이버시권> 보고서에서 국가와 기업에 대하여 “인공지능 시스템의 설계, 개발, 배치, 판매, 구입, 운영의 수명 주기 전반에 걸쳐 체계적으로 인권실사를 수행”할 것을 권고하고, “그 인권실사의 핵심 요소

79) UNITED NATIONS(2018). para.53; 55; 62; 63; 68.

80) Office of the High Commissioner for Human Rights(2020). Identifying and Assessing Human Rights Risks related to End-Use. <https://www.ohchr.org/Documents/Issues/Business/B-Tech/identifying-human-rights-risks.pdf(접근일: 2022. 8. 15.)>; 김기중 외(2021), 부록 I.

는 정례적이고 포괄적인 인권영향평가여야 한다" 고 강조하였다.[81]

유엔인권이사회는 최근의 결의에서 인공지능에 대한 인권실사를 강하게 요구하였다. 2021년 10월 13일 <디지털시대 프라이버시권 결의>[82]는 "감시, 인공지능, 자동화된 의사결정 및 기계 학습 분야에서 개발된 기술, 얼굴 인식 및 감정 인식을 비롯한 프로파일링, 추적 및 생체 인식 분야에서 개발된 기술 등 신기술(new and emerging technologies)이 적절한 보호 장치 없이 프라이버시 권리 및 표현의 자유와 간섭 없이 의견을 가질 권리, 평화적 집회 및 결사의 자유를 비롯한 여타의 인권의 향유에 미치는 영향이 증가하고 있음을 상기" 하면서, "인공지능 등 신기술의 계획, 설계, 개발 및 배치에 있어, 해당하는 국제인권법에 따른 의무를 준수하여 적정한 규제 및 기타 적절한 제도를 도입함으로써 프라이버시권 및 여타 인권에 미치는 위험을 최소화할 수 있고 최소화하여야" 한다고 지적하였다. 더불어 이는 "인권에 미치는 부정적인 영향을 평가, 방지, 완화하는 인권실사를 시행하고, 인간의 감독 및 시정 체계를 수립함으로써 이루어질 수 있다" 고 보았다. 이에 인권이사회는 "국가 및 해당되는 기업이 설계, 개발, 배치, 판매 또는 구입 및 운영하는 인공지능 시스템의 수명주기 전반에 걸쳐 인권실사를 실시할 것을 권장" 하였다.

2. 유럽평의회 권고와 인권·민주주의·법치 영향평가

가. 유럽평의회 권고

유럽평의회 인권위원장은 2019년 <인공지능 블랙박스 개봉: 인권 보호를 위한 10단계>[83]에 대한 권고를 발표하면서 인공지능이 인권에 미치는 부정적인 영향을 예방하고 완화하기 위한 첫번째 방안으로 인권영향평가를 권고하였다. 국가는 공공기관이 구입, 개발 또는 배치하였거나 예정하고 있는 인공지능 시스템에 대하여 인권영향평가 실시를

81) UNITED NATIONS(2021), para.20; 56; 60(a).
82) Resolution adopted by the Human Rights Council on 7 October 2021, 48/4, Right to privacy in the digital age, UN Doc. A/HRC/RES/48/4(13 October 2021).
83) Council of Europe Commissioner for Human Rights(2019); 김기중 외(2021), 부록 III.

요구하는 법률과 규제 조치를 취해야 한다. 이때 인권영향평가는 독립된 감독기구나 관련 전문성을 갖춘 외부기관의 검토를 포함해야 하며, 공공기관은 유의미한 외부 검토기관으로서 국가인권기구를 고려해야 한다. 인권영향평가의 적용을 받지 않았거나, 인권영향평가가 인공지능 시스템의 실제적인 인권 침해 위험을 나타냈음에도 식별된 위험을 방지하거나 완화하기 위한 조치, 안전장치 또는 방법을 채택하지 않은 상황에서는 인권을 간섭할 가능성이 있는 인공지능 시스템을 배치 및 사용하지 말아야 한다.

유럽평의회는 2020년 4월 8일 알고리즘 시스템의 인권영향에 대한 각료위원회 권고 CM/Rec(2020)1[84]를 채택하였다. 이 권고와 부록 <알고리즘 시스템의 인권영향에 대한 대응 지침>은 국가에 대한 일반 요구로, 인권 침해를 예방, 탐지, 금지 및 구제하는 효과적이고 예측 가능한 입법을 추진하고, 인권영향평가 등 지속적인 검토 체계를 마련할 것을 요구하였다. 더불어, 알고리즘 시스템의 위험에 대응하는 민주적 참여 및 인식을 제고하고, 규제 및 감독을 위한 제도적 체계를 수립해야 한다. 특히 국가는 알고리즘 시스템의 전체 수명 주기 동안 개별 시스템의 인권영향 및 다른 기술과의 상호 작용을 정기적으로 평가해야 한다. 평가는 영향을 받거나 영향을 받을 가능성이 있는 사람들과의 광범위하고 효과적인 협의를 기반으로 수행되어야 한다. 이때 국가는 독립적 감독 기관, 평등기구, 국가인권기구, 대학, 표준 수립 기구, 서비스 운영자, 알고리즘 시스템 개발자 및 특히 인권옹호자 등 다양한 분야의 관련 비정부 기구와 긴밀하게 협력해야 한다.

유럽평의회는 특히 인권영향평가를 국가가 의무적으로 취하여야 할 예방적 조치로 보고 상세한 요구사항을 설명하였다. 인권영향평가는 인공지능 수명 주기의 모든 단계에서 잠재적 위험을 평가하고 그러한 위험을 예방하거나 완화하기 위한 조치, 보호장치 및 메커니즘을 수립하여야 한다. 인권에 높은 위험을 수반하는 모든 알고리즘 시스템에 대하여 인권영향평가가 의무화되어야 한다. 또한 국가는 공공조달 이전, 개발기간 중, 정규 일정 및 상황별 배치 전반에 걸쳐, 정기적이고 협의적으로 인권영향평가를 수행해야 한다. 고위험 알고리즘 시스템과 관련된 모든 인권영향평가가 독립적인 전문가 검토 및 검사를 위해 제출될 수 있어야 하고, 국가에 의해 또는 국가를 위해 수행되는 인권영향평가는 공개되어야 한다. 후속 조치로는 동적 검사 방법 및 출시 전 시험을 수행하고, 영향을 받을 수 있는 개인과 집단 및 관련 현장 전문가와 협의하여야 한다. 인권영향평가

84) Council of Europe(2020); 김기중 외(2021), 부록 IV.

가 완화할 수 없는 중대한 인권 위험을 식별하는 상황에서는 공공기관이 해당 알고리즘 시스템을 구현하거나 사용해서는 안 된다. 이미 배치된 알고리즘 시스템과 관련하여 위험이 식별되면 최소한 위험 완화를 위한 적절한 조치가 취해질 때까지 시스템 구현을 중단해야 한다. 식별된 인권 침해는 즉시 해결 및 구제되어야 하며 추가 침해를 방지하기 위한 조치가 취해져야 한다.

민간부문 역시 예방적 조치로서 인권영향평가를 실시하면서 영향을 받는 개인 및 집단의 적극적인 참여를 포함하고 가능한 한 공개적으로 수행하여야 한다. 인권영향평가의 후속 조치로서, 식별된 오류를 가능한한 신속하게 해결하고, 적절한 경우 관련 활동을 일시 중단시켜야 한다. 이를 위해서는 알고리즘 시스템의 설계, 검사 및 배치 단계 전반에 걸쳐 정기적이고 지속적인 품질 보증 검사와 실시간 감사가 필요하다. 영향을 받는 개인과의 정기적인 협의가 추가적으로 필요하며, 이는 부정적인 인권영향을 악화시키고 고착화할 수 있는 피드백 순환구조의 위험을 고려할 때 특히 중요하다.

나. 인공지능 시스템에 대한 인권・민주주의・법치 영향평가

2020년 12월 유럽평의회의 인공지능 특별위원회(Ad hoc Committee on Artificial Intelligence, 이하 ‘CAHAI’)는 <인공지능의 법적 프레임워크에 대한 타당성 연구>를 채택하였다.[85] 이 보고서 9장은 인권, 법치 및 민주주의에 관한 유럽평의회의 표준에 기반하여 인공지능의 법적 프레임워크의 준수를 보장하기 위해 필요한 실질적인 후속 메커니즘에 대한 내용을 담고 있다. CAHAI는 그러한 메커니즘의 하나인 인권영향평가를 상세히 검토하였으며, 2021년 5월 21일 <인공지능 시스템에 대한 인권・민주주의・법치 영향평가> 초안 문서를 발표하였다. 이 문서는 기존의 영향평가 모델에 대한 검토를 바탕으로 인공지능에 대한 인권・민주주의・법치 영향평가(Human Rights, Democracy, and Rule of Law Impact Assessment of AI, 이하 ‘HRDRIA’) 모델을 제시하고 있다.

85) Ad hoc Committee on Artificial Intelligence(2020).

1) 인공지능 영향평가의 주요 특성

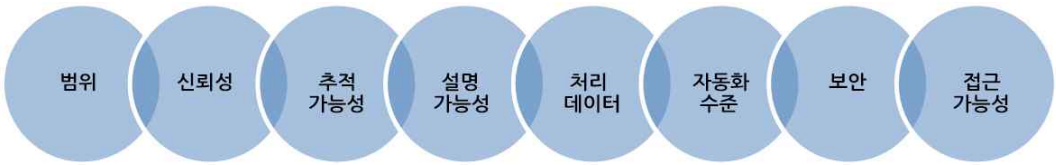
CAHAI는 기존의 인권영향평가의 관행과 경험에 기반해야 한다면서도 인공지능에 대한 HRDRIA를 수행할 때 고려해야 할 특성을 검토하고 있다. 우선 HRDRIA는 공공 및 민간 모두에 적용됨을 밝히고 있다. 인권 맥락에서 공공 및 민간 행위자의 역할과 의무의 범위가 다르기는 하지만, HRDRIA를 수행하는 것은 모든 행위자에게 중요하고, 또한 인공지능과 관련하여 공공 및 민간의 구별이 모호한 경우가 많기 때문이다.

CAHAI는 일반적인 인권영향평가와 다른, 인공지능 시스템의 평가 변수로서의 주요 특성을 다음과 같이 제시하고 있다.

첫째, 인공지능 시스템이 운영되는 지정학적, 사회적, 경제적 맥락이다. 동일한 인공지능 시스템이라도 사용되는 맥락에 따라 위험성이 달라질 수 있기 때문이다. 이러한 맥락을 파악할 때, 설계자, 개발자, 혹은 고객이 요구하는, 시스템의 목적을 고려할 필요가 있다. 예를 들어, 금융 거래를 추적하는 인공지능 시스템이 돈세탁 흐름을 파악하기 위한 목적으로 사용될 수도 있고, 기업이나 개인의 현금 자산의 흐름을 파악하기 위한 목적으로 사용될 수도 있다. 인공지능 시스템의 '사용자'가 누구인지 역시 마찬가지로 중요하다.

둘째, 인공지능 시스템의 기반 기술도 영향평가의 변수로 고려해야 한다. 인공지능 기술은 계속 진화하고 있으며, 각 기술의 차이와 함의를 이해하는 것이 중요하다. 이때 범위, 신뢰성, 추적가능성, 설명가능성, 사용된 데이터, 자동화 수준, 보안, 접근가능성 등 잠재적 위험과 관련된 인공지능 기술의 8가지 차원을 고려해야 한다.

[그림 9] CAHAI 잠재적 위험과 관련된 인공지능 기술의 8가지 차원



‘범위’는 기반 기술의 범위 내에서 인공지능 시스템이 사용되는가의 문제이다. 인공지능 시스템은 데이터와 알고리즘에 따라 성능(오류율과 같은)이 달라지는데, 학습 방법과 알고리즘의 형태를 아는 것은 인공지능 시스템의 결과를 해석하는 것뿐 아니라, 데이터 조달 및 가지치기(pruning), 잠재적 편향의 검토와도 관련된다. 예를 들어, 재판 중에 생성된 총계 형태의 포렌식 증거를 특정 개인의 유죄 여부 가능성의 근거로 사용하는 것은 범위를 벗어난 사례에 해당한다. 인공지능 시스템 결과물의 일관성은 기반 기술에 따라 달라지는데, ‘신뢰성’의 측정은 위험평가에 유용하다. 예컨대, 번역 인공지능의 약간의 부정확성은 감내할 수 있지만, 자율주행 자동차에서는 그렇지 않다. ‘추적가능성’은 인공지능 시스템이 왜 특정한 행위를 하는지 설명할 수 있도록 설계될 것을 요구한다. 추적가능성의 수준은 ‘설명가능성’과 연결된다. 그러나 설명가능성은 소통의 측면도 수반하는데, 전문가에게는 설명가능하더라도 일반 공중은 이해하기 힘들 수도 있기 때문이다. ‘처리 데이터’와 관련해서는 민감한 데이터가 더 위험하며, 데이터의 규모, 데이터셋의 다양성, 비차별성 등의 특성에 따라 위험성이 달라진다. ‘자동화 정도’는 기술적 측면과 밀접한 관계 속에서 검토되어야 한다. 인간 운영자가 있을 경우 완전 자동화 시스템에 비해 인간 오류의 가능성이 높아지는데, 항상 그런 것은 아니다. 설령 인간 운영자가 관여하더라도 운영자가 시스템의 잘못된 판단을 수정하는 것이 아니라 인공지능 시스템의 결과를 무조건 수용할 수 있기 때문이다. ‘보안’ 위험이 크면 당연히 해악의 위험이 커지며, 기술의 ‘접근가능성’은 사회적 정보격차와 관련된다.

셋째, 인공지능 시스템에 관련된 행위자와 개발 단계가 영향평가의 변수로 고려되어야 한다. HRDRIA는 인공지능 시스템의 전체 생애주기에서 지속적인 평가도구가 되어야 하는데, 생애주기에는 설계자, 개발자, 배포자, 운영자, 사용자 등 여러 행위자가 서로 다른 역할을 하게 된다. 인공지능 시스템 개발에도 서로 다른 일련의 절차가 포함된다. 예를 들어, 구성(configuration), 자동화, 데이터 수집 및 검증, 특성 엔지니어링, 테스트 및 디버깅, 자원 관리, 모델 분석, 절차 관리, 메타데이터 관리, 기반시설 제공, 모니터링 등이다. 어떤 단계에서 평가를 하느냐에 따라 영향평가의 내용이 달라질 수 있다.

넷째는 이해관계자의 참여가 성공적인 HRDRIA를 위해 필수적이라는 점이다. 따라서 관련된 이해관계자를 식별할 수 있는 효과적인 메커니즘이 필요하다.

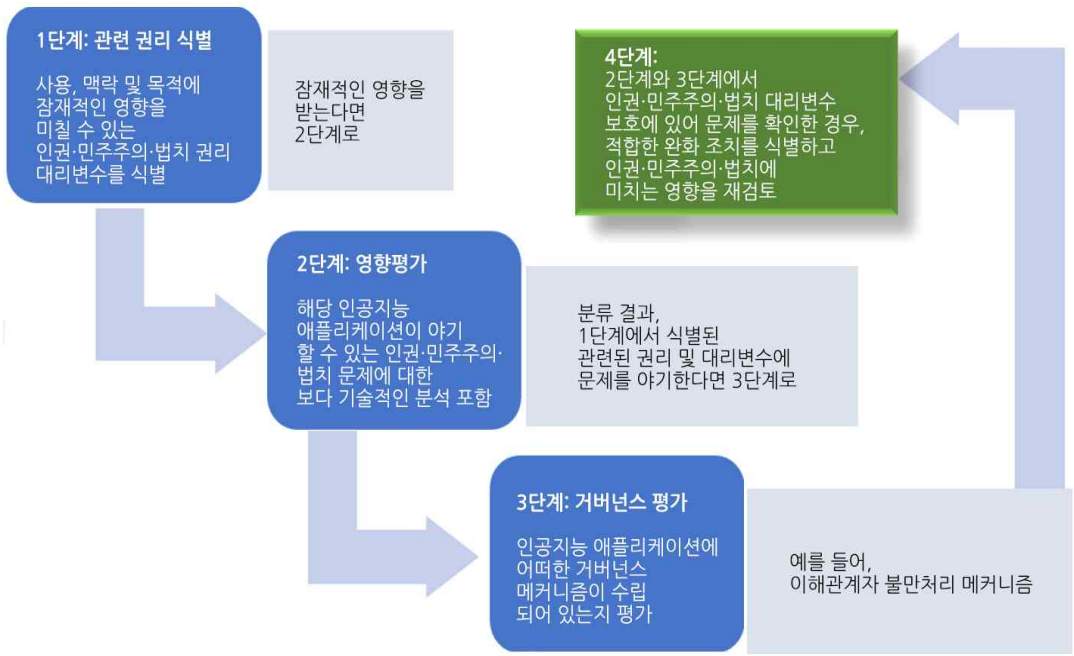
한편, CAHAI는 인공지능이 인권에 미치는 영향과 더불어, 민주주의 및 법치에 대한 영

향을 평가하는 것이 쉽지 않다는 점을 지적한다. 민주주의와 법치 모두 그 의미의 폭이 매우 넓기 때문에 평가의 기준으로 사용하기 힘들기 때문이다. 이 때문에, CAHAI는 민주주의 및 법치와 직접적으로 관련된 인권을 대리변수(proxy)로 사용하는 방법을 제안한다. 예를 들어, 집회 결사의 자유는 민주주의 수준을 평가할 수 있는 하나의 지표이므로, 집회 결사의 자유라는 인권에 미치는 영향을 민주주의에 대한 영향을 나타내는 대체변수로 사용할 수 있다는 것이다.

2) HRDRIA 수행 절차

CAHAI는 현재의 인공지능 인권영향평가의 경험에 기반하여, 인공지능의 HRDRIA 수행 모델을 제안한다. 우선 HRDRIA를 언제 수행해야 하는지와 관련하여, CAHAI는 '초기 평가'에서 인권 위험이 높게 나왔을 경우 수행할 것을 제안하고 있다. 모든 인공지능 시스템을 대상으로 수행하는 것은 시간과 비용이 과도하게 들기 때문이다.

[그림 10] CAHAI 인공지능에 대한 인권・민주주의・법치 영향평가 절차



따라서 인권에 부정적 영향을 미칠 위험이 큰 일부 시스템에 대해서만 HRDRIA를 수행한다면, 이러한 초기 평가의 기준을 개발할 필요가 있다. 개인정보보호 영향평가의 경험에 비추어볼 때, 이러한 기준으로는 애플리케이션의 규모, 형태와 목적, 체계적으로 인간과 상호작용하는 정도, 특별히 위험한 사용 사례(예를 들어 얼굴인식, 딥페이크, 소셜네트워크) 등이 고려될 수 있을 것이다.

CAHAI가 제안하는 HRDRIA의 방법론은 다음과 같다.

1단계는 인공지능 시스템에 의해 부정적 영향을 받을 가능성이 있는 관련 권리를 파악하는 것이다.

2단계에서는 해당 권리에 대한 영향평가를 시행한다. 이 영향평가는 기술적, 비기술적 측면을 모두 포함한다. 첫째, 기술적 분석은 인공지능 애플리케이션에 적용된 기반 기술 및 부정적 영향을 막기 위한 기술적 특성, 예를 들어 설명가능성, 투명성, 사이버보안, 의도된 사용을 넘어선 활용을 막기위한 보호조치 등에 초점을 맞춘다. 둘째, 비기술적 분석은 시스템이 작동하는 사회-기술적 환경, 인공지능 애플리케이션의 보급 및 사용에 필요한 역량과 기술을 분석한다. 또한, 전반적인 가치 사슬 및 생애주기에서 인공지능 보급의 위험성에 대한 식별, 해결, 추적 또한 검토되어야 한다. 셋째, 다른 인공지능 인권영향평가와 다르게 HRDRIA는 민주주의 및 법치의 대체변수에 해당하는 기본권에 대한 영향 분석을 포함한다.

3단계는 거버넌스 메커니즘에 대한 평가이다. 잠재적 위험을 완화하는데 도움이 될 거버넌스 메커니즘이 있는지 검토한다. 예를 들어, 이해관계자 참여나 불만처리 메커니즘을 검토해야 한다. 불만처리 메커니즘은 HRDRIA의 지속적인 학습 과정에서 필수요소이다.

4단계는 거버넌스 메커니즘과 다른 완화 조치들이 부정적 영향을 완화하는 해결책을 제공하는지에 대해 지속적으로 점검(evaluation)하는 것이다. 또한 영향평가는 인공지능 시스템, 해당 시스템의 환경, 거버넌스 메커니즘의 변화를 고려하여 지속적으로 이루어져야 한다.

3) 추가적인 고려사항

인공지능 영역에서는 기본권 사이에 충돌이 종종 발생할 수 있으므로, 이를 어떻게 다뤄야할 것인지에 대한 가이드가 필요하다. 또한, 인공지능이 공공 분야에 적용될 경우 추가적인 안전조치가 필요한지 고려할 필요가 있다.

HRDRIA는 부정적 영향 완화를 위한 지속적인 학습 과정이다. HRDRIA는 특정 시점에서의 평가를 제공하지만, 일회성 행사가 아니라 반복되어야 한다. 완화조치 적용의 결과 확인을 위해, 혹은 새로운 위험이 나타날 경우 HRDRIA는 반복될 수 있다.

HRDRIA는 개발자 뿐만 아니라 판매자, 조달자, 배포자에 의해서도 수행되어야 한다. 또한 HRDRIA는 이해관계자의 참여를 강조한다. 즉, 인공지능의 기술적 측면이나 개발 조직 뿐만 아니라, 내부 및 외부의 이해관계자의 참여에도 초점을 맞추어야 한다. 참여의 방식은 인권에 미치는 영향의 심각성, 규모, 회복불가능성 등에 따라 달라질 수 있다. HRDRIA의 전 과정에서 이해관계자가 참여할 수 있어야 하며, 이해관계자가 이해할 수 있도록 정보를 제공해야 한다.

부정적 영향이 발견될 경우 구제조치에 대한 접근이 중요하게 고려되어야 한다. 이와 관련하여 운영 단계에서는 불만처리 메커니즘을 갖추는 것이 중요하다.

3. 덴마크 디지털활동 인권영향평가

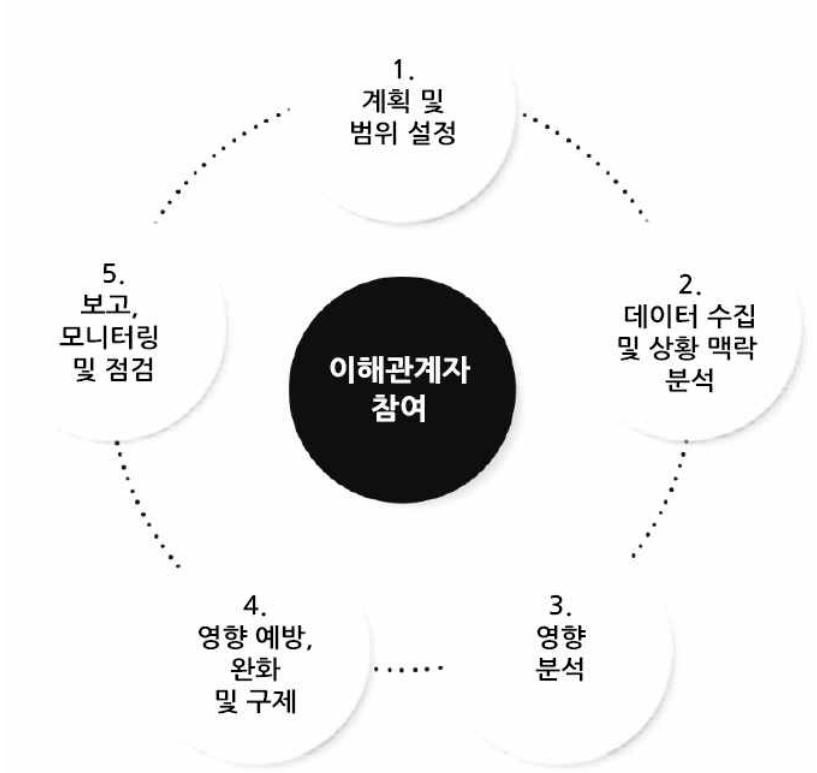
덴마크 국가인권기구는 2020년 <디지털활동 인권영향평가 지침(Guidance on Human Rights Impact Assessment of Digital Activities, 이하 '덴마크 인권영향평가')>[86]를 발간하였다. 이 지침은 인공지능 등 디지털활동, 제품 및 서비스가 야기, 기여, 관련된 위험의 특성을 평가하고 해결하기 위한 지침이 현재 부족하고 관련 인권영향평가 또한 신생 분야라고 지적하며, 그럴수록 이해관계자들과 많이 대화하도록 하는 인권기반 접근법을 적용하고 인권영향평가를 실시하면 얻는 바가 있을 것이라고 설명하였다.

덴마크 국가인권기구는 디지털활동 인권영향평가에서 고려해야 할 필수 핵심 요소를

86) The Danish Institute for Human Rights(2020b).

△국제 인권 기준 △인권 기반 절차 △책임성으로 꼽았다. 즉, 영향평가의 기준은 국제 인권기준으로 구성되어야 하고, 평가 절차 자체에서 차별금지, 권리주체 참여 및 투명성 등 인권 원칙을 존중하여야 하며, 의무주체가 인권 침해에 대하여 계속적인 책임을 가져야 한다는 것이다.

[그림 11] 덴마크 디지털활동 인권영향평가 절차



가. 평가의 절차

덴마크 인권영향평가는 인권실사가 기업 활동 전반에 걸쳐 시행되어야 하는 반복적인 과정이라는 점에서 인권영향평가 또한 언제, 어떻게 실시할지는 사업별로 특유하다고 설명한다. 얼마나 정밀한 평가를 수행해야 하는지 역시 규격화할 수 없으며, 식별된 위험, 그 심각도, 잠재적 또는 실제 영향에 대한 회사의 자원 및 참여, 기타 다양한 요인에 따

라 달라진다.

인권영향평가를 실시한 후로도 해당 제품을 새롭거나 위험이 높은 시장에 출시할 때, 이용약관의 중대한 변경이 있을 때, 특정 시장에서 제품을 철수하는 등 디지털활동, 제품 및 서비스의 규모, 범위, 사용 또는 적용이 변경될 때에는 재평가되어야 한다. 법률, 규정 또는 시장에 대한 중요한 변화가 있거나 사회적, 정치적 상황에 중대한 변화가 있을 때에도 최초 인권영향평가의 결과를 재평가해야 한다.

덴마크 인권영향평가는 다음의 다섯 단계를 거치며 수행된다.

- 1단계: 계획 및 범위 설정
- 2단계: 데이터 수집 및 상황 맥락 분석
- 3단계: 영향 분석
- 4단계: 영향 예방, 완화 및 구제
- 5단계: 보고, 모니터링 및 점검

이때 평가가 반복적인 과정임을 인식하고 평가 내내 지속적인 학습과 분석을 촉진할 필요가 있다.

○ 1단계 : 계획 및 범위 설정

인권영향평가가 효과적으로 수행되고 원하는 결과를 달성하려면 올바른 계획과 범위 설정이 필수적이다. 범위 설정의 목표는 디지털활동의 유형, 인권적 맥락, 관련 이해관계자, 평가 결과의 활용처와 관련한 정보를 검토하여 인권영향평가의 매개변수를 정의하려는 데 있다. 수집된 정보는 평가를 위한 임무정의서(Terms of Reference, TOR)을 작성하고 맥락 분석을 시행하며 차후에 영향을 분석하는 데 사용되지만, 범위 설정이나 TOR은 이후 평가에 소요되는 시간, 새롭게 발견되는 이슈나 인권영향에 따라 유동적으로 변경될 수 있다. 기관과 담당자는 이 단계에서 피평가기관으로부터 독립적인 인권영향평가팀의 구성과 이해관계자의 참여에 대한 결정을 내리게 된다.

덴마크 인권영향평가는 디지털활동 인권영향평가 TOR 작성을 위한 체크리스트를 다음과 같이 제시하였다.

[표 5] 덴마크 디지털활동 인권영향평가 임무정의서(TOR) 체크리스트

과업 요소	질의 예시
배경 요소	• 개발 및 출시 단계, 애플리케이션 유형, 애플리케이션 용도, 위치, 부문, 규모 등에 대한 정보를 비롯하여 평가 대상 디지털 프로젝트, 제품 또는 서비스가 명확하게 서술되어 있는가? • 인권영향평가의 근거, 즉 인권영향평가의 내적/외적 동인이 서술되어 있는가?
과제 서술	• 인권영향평가의 목적과 의도하는 결과가 명확하게 표현되어 있는가? 제한사항을 언급하고 있는가?(개발 초기 단계로 인해 권리주체 특정/참여의 어려움, 인권영향평가팀이 사용할 수 없는 회사 데이터 등) • 인권영향평가에서 고려해야 하는 관련 배경 정보가 과제 서술에 포함되어 있는가?(다른 인권실사 활동, 개인정보보호 영향평가, 윤리적 영향평가 및 기타 평가를 실시한 결과 등) • 인권영향평가의 범위가 명확하게 정의되어 있고 그 범위가 평가할 인권영향을 포괄적으로 다루는가?(디지털 프로젝트, 제품 또는 서비스가 야기하거나 기여하는 실제적 및 잠재적 영향, 직접적으로 관련된 영향, 누적적 영향 등) • 인권영향평가의 목적상 부정적인 영향을 식별하는 데 우선순위를 두었을 때 고려되어야 할 영향의 범위가 부정적인 영향과 긍정적인 영향을 명확하게 구분하는가?
방법론	• 국제인권기준이 평가의 기준으로 명확하게 명시되어 있는가? 특정 권리주체의 권리(아동 권리 등)가 [평가에] 요구되는 인권 기준에 포함되어 있는가? • 법률 및 표준, 회사 및 금융기관의 기준 및 요구사항이 명확한 참조사항으로 고려되는가?(관련 디지털 프로젝트, 제품 및 서비스의 개발, 사용 및 판매에 대한 회사 고유 윤리 원칙 등) • 인권 기반 접근 방식(참여, 차별금지, 역량 강화, 투명성 및 책무성 원칙 등)의 적용이 인권영향평가의 작업 방법론에 필수적인 내용으로 명확히 명시되어 있는가? • 영향 심각도 평가를 위한 지표에서 규모, 범위 및 회복불가능성이 명확히 포함되어 있는가? • 방법론 요구사항은 적용할 완화 계층이 국제 인권 기준 및 원칙에 합치되어야 함을 설명하고 있는가? • 방법론이 포괄적인 이해관계자(인권영향평가에 참여할 것으로 일반적으로 식별되고 설명되는 권리주체, 의무주체 및 기타 관련 이해관계자) 참여를 명확히 예정하고 있는가? 또한 독립적인 인권 전문가 및 기타 인권 관계자가 인권영향평가 목적을 위한 관련 이해관계자로 설명되어 있는가?

	• 방법론이 포괄적인가? 즉, 범위 설정, 데이터 수집 및 맥락 분석, 영향 분석 및 평가, 완화 조치 개발, 모니터링 단계 및 보고가 포함되는가? • 방법론은 인권영향평가를 완수하기 위해 문헌 조사와 (권리주체 및 그 대리인, 의무주체 및 기타 관련 당사자 참여를 통하여) 직접적인 데이터 수집을 모두 명확히 요구하는가? • 알려진 제한 사항이 처음부터 명확하게 명시되어 있는가? 제한이 위에 명시된 요소들을 방해하는 경우 그러한 제한이 정당화되는가?
요구되는 전문성	• 인권영향평가팀의 기량과 경험에 대한 세부 정보가 제공되고 있는가?(인권 및 기타 요구되는 전문 지식, 기술적 지식, 성인지, 언어 능력, 현장 지식 등) • 필요한 경우 통역사와 현장 담당자의 참여를 보장하는 규정이 마련되어 있는가?
거버넌스 및 보고 구조	• 인권영향평가의 거버넌스 구조가 명확하게 설명되어 있는가?(인권영향평가팀의 역할 및 독립성, 회사 담당자 및 상대방의 역할, 자문 위원 또는 동료 검토 메커니즘의 역할 등) • 인권영향평가 보고서(전체 또는 일부)의 발간을 비롯한 보고 요구사항이 명확히 규정되어 있는가? 또한 영향평가 결과에 관하여 권리주체 및 기타 이해관계자에게 공유하는 다른 방식도 규정되어 있는가? • 인권영향평가 및 관련 후속 활동에 대한 고위 경영진/임원의 역할과 내부 관리구조가 명확하게 설명되어 있는가?(인권영향평가팀 또는 다른 사람이 결과를 어떻게 보고하고, 인권영향평가의 정보가 회사에 어떻게 회람되며, 권고사항의 이행에 대한 책임이 누구에게 있는지 등)
업무 계획, 일정표 및 예산	• 중간 및 최종 결과물 등 인권영향평가 업무 계획이 명확하게 설명되어 있는가? • 인권영향평가 예산이 명확하고 지정된 평가를 수행하기에 충분한가? 특히 해당 예산으로 인권영향에 대한 유의미한 평가를 수행하는데 필요한 데이터 수집이 가능한가? • 인권영향평가의 기간이 특정되어 있으며 평가 완수에 필요한 연구와 이해관계자 참여에 충분한 시간을 보장하는가?

○ 2단계 : 데이터 수집 및 상황 맥락 분석

2단계에서는 사용자, 잠재적 및 실제적으로 영향을 받는 권리주체, 특히 취약 집단의 인권 향유에 대한 기본 데이터를 이해관계자로부터 수집한다. 앞서 범위 설정 단계는 주로 탁상 연구 및 2차 소스 분석에 의존하지만 이 단계에서는 1차적인 데이터를 수집하고 대면 또는 가상 온라인으로 인터뷰하는 등 이해관계자의 참여와 협의를 중시한다.

어떤 인권영향평가 문헌과 방법은 이 단계를 ‘증거 수집’ 단계라고 지칭하기도 하

며, 다른 영향평가 분야에서는 ‘기준 연구’ 또는 ‘기준 개발’이라고 부르기도 한다. 유엔 인권최고대표실은 데이터 수집에 있어 인권기반접근법은 ①참여, ②데이터 세분화, ③자기 확인, ④투명성, ⑤개인정보보호, ⑥책무성의 6가지 측면을 살필 것을 권고한 바 있다.[87] 덴마크 인권영향평가는 디지털활동에 대한 데이터 수집에 있어서도 인권기반접근법의 기준을 다음과 같이 적용할 것을 제안하였다.

[표 6] 데이터 수집의 인권기반접근법

인권기반접근법	설명	디지털활동 적용 예시
참여	데이터 수집 과정에 관련 이해관계자 및 권리주체(대리인 또는 대표자의 형태일 수 있음)가 포함되어야 한다. 실무적으로 이는 인권영향평가팀이 성인지적 접근 방식을 취하고 여성, 아동, 장애인, 노인, 성소수자, 이주민, 난민 등 취약하고 소외될 수 있는 개인 및 집단에 특별히 주목해야 한다는 의미이다.	‘스마트 채용 시스템’이 여성에게 영향을 미치는 차별적 결과를 초래할 수 있음이 식별되면, 여성 권리주체 및 여성단체가 데이터 수집 과정에 직접 참여해야 한다.
데이터 세분화	데이터 세분화를 통해 연구자는 다양한 인구집단에 미치는 불평등 영향을 비교할 수 있다. 단순한 데이터 평균은 이면의 차이를 은폐할 수 있다. 이와 대조적으로, 세분화된 데이터는 집단 간에 차별적으로 미치는 인권영향을 보여줄 수 있다.	데이터 수집 결과 알고리즘 신용 위험 점수를 받은 사람들 중 극히 일부만이 차별을 받았다고 생각한다면 문제의 범위가 작음을 시사한다. 이와 대조적으로, 세분화된 데이터는 고충 신고 대부분이 어느 한 소수인종에서 나왔다는 사실을 보여줄 수 있다.
자기 확인	“해를 끼치지 말라”는 중요 원칙대로, 데이터 수집은 참가자에게	성소수자 개인에 대한 온라인 학대가 잠재적인

87) The Office of the United Nations High Commissioner for Human Rights(2018), ”A Human Rights-Based Approach to Data“. <https://www.ohchr.org/Documents/Issues/HRIndicators/GuidanceNoteonApproachtoData.pdf(접근일: 2022. 9. 10)>.

	부정적인 영향을 미치지 않아야 한다. 참가자는 자신의 정체성을 자유롭게 정의할 수 있는 선택권과 자신의 특성에 대한 정보를 공개할지 여부를 선택할 수 있는 권한을 부여받아야 한다.	영향으로 식별된 경우, 권리주체 집단의 개인에 대한 데이터를 직접 수집할 때 당사자가 자신의 특성에 대한 정보를 공개할지 여부를 수집 과정에서 결정할 수 있어야 한다.
투명성	인권영향평가팀은 인권영향평가에서 사용된 방법론과 목적을 비롯한 평가 절차에 대해 명확히 해야 한다. 여기에는 데이터 수집 과정 그 자체에 대한 투명성도 포함된다(어떤 이해관계자가 참여하고 그 선정 방식은 무엇인지 등). 인권영향평가팀이 무엇을 할 것이고 달성하거나 약속할 수 없는 사항이 무엇인지 또한 분명해야 한다.	인권영향평가에서 온라인 설문조사를 사용하는 경우, 모든 설문 응답자에 대하여 무엇에 참여하고 있는지, 응답이 어떻게 처리되는지, 사생활의 권리가 어떻게 보호되는지, 어떤 결과가 나타날 수 있는지 등을 분명히 밝혀야 한다.
사생활	수집된 데이터는 기밀로 유지되어야 하며 연구자는 공개하거나 사용하는 데이터에서 개별 참가자를 식별할 수 없도록 보장해야 한다. 이는 민감한 주제를 다루고 참가자가 보복 위험에 직면할 수 있는 인권영향평가의 경우에 특히 중요하다. 따라서 연구자는 참가자의 개인정보는 물론 응답에 대해서도 강력한 개인정보보호 조치를 취해야 한다.	디지털활동 인권영향평가에서 참가자 개인정보보호를 강화하기 위해 통신을 암호화할 수 있는 온라인 참여 형식을 이용할 수 있다.
책무성	인권영향평가의 데이터 수집 단계에서 수집된 정보는 의무주체(가장 두드러진 국가 및 기업 행위자)가 인권에 미치는 영향에 대해 책임을 지는 데 사용되어야 한다. 데이터를 수집하는 연구자 역시 데이터의 품질과 신뢰성에 대하여 책임을 져야 한다.	

이렇게 수집된 데이터를 통해 평가팀은 인권 향유의 현재 상태에 대한 맥락 분석을 수행할 수 있다. 맥락 분석은 인권에 미치는 실제 영향을 식별하고 향후 영향을 더 잘 예측하는 데 도움이 된다.

이 단계에서는 인권의 특정 측면에 주목하여 데이터를 수집해야 할 뿐 아니라, 후속 영향 예방, 완화 및 개선에 주목해야 한다. 이를 위하여 인권영향평가팀은 구조적, 절차적, 결과적 수준에서 질적 및 양적 지표를 모두 사용해야 한다.

유엔 기업과 인권 이행지침에 따르면, 기업은 부정적인 인권영향이 해결되었는지 여부를 확인하기 위해 자신들의 대응 효과를 추적해야 한다. 이 추적은 적절한 질적 및 양적 지표(indicators)를 기반으로 해야 한다. 다만 이들 지표가 인권영향평가에 유용한 도구임에 분명하지만, 인권영향 분석은 항상 질적이고 설명에 기반한 분석을 요구하기 때문에 지표 및 기타 '측정값'에만 의존할 수 없다. 지표는 강력한 질적 차원으로 평가에 가치를 더하는 도구일 뿐, 이를 대체할 수 없다는 점을 염두에 두어야 한다. 결국 인권영향평가 실무자에게 지표는 소위 '위험 신호'로서 도움이 될 수 있으나, 실제적 및 잠재적인 인권영향을 완전히 이해하기 위해서는 관련 권리주체, 의무주체 및 기타 관련 당사자와 협의하는 등 질적 방법을 사용하는 조사가 추가적으로 이루어져야 한다.

○ 3단계 : 영향 분석

부정적인 인권영향은 작위 또는 부작위가 개인 및 집단이 인권을 향유할 수 있는 능력을 전체 또는 부분적으로 제한할 때 발생한다. 3단계에서는 실제적 또는 잠재적인 인권영향을 식별하고 심각도를 평가하기 위해 앞서 수집된 데이터를 분석한다. 분석에는 국제 인권 기준, 사업 비교, 이해관계자 참여 결과 등을 활용한다.

인권영향 분석은 '즉각적'으로 보이는 영향뿐 아니라 사업이 야기하고 기여했거나 그럴 수 있는 모든 영향, 직접적으로 관련된 영향을 고려하여야 한다. 영향 분석에는 영향의 범위, 규모 및 회복 불가능성을 고려하여 영향의 '심각도'를 평가하는 과정이 포함되어야 한다. 인권영향의 심각도가 높을수록 신속하고 엄격한 조치가 필요하다. 이때 영향을 경험하였거나 경험할 수 있는 사람의 관점을 고려해야 한다.

무엇보다 인권영향평가가 인권 존중에 기여하기 위해서는 부정적인 인권영향을 우선적으로 식별하고 해결하는 데 중점을 두어야 한다. 긍정적인 영향이 나타날 수 있지만

‘긍정적인’ 인권영향을 식별하는 것은 주요 목표가 아니며 부정적인 영향을 식별하고 해결하는 데 방해가 되어서는 안 된다.

○ 4단계 : 영향 예방, 완화 및 구제

효과적인 영향 예방, 완화 및 개선(영향 관리)을 위한 계획은 인권영향평가의 필수적인 부분이다. 이 단계에서 기관, 평가팀 및 이해관계자는 협력하여 부정적인 인권영향을 예방, 완화 및 개선하기 위한 계획을 수립한다. 모든 인권영향을 해결하는 것이 목표이며 가장 심각한 영향을 우선적으로 고려해야 한다. 권리주체 및 그 대리인은 영향 예방, 완화 및 개선 조치를 계획, 시행 및 모니터링하는 데 유의미하게 참여해야 한다.

식별된 영향에 대한 조치는 주로 부정적인 인권영향을 방지하고 감소시키는 데 중점을 두어야 한다. 이때 여러 영향평가 방법론에서 완화 조치에 대하여 대부분 다음과 같은 계층적 접근 방식(Mitigation Hierarchy)을 취하고 있음을 참고할 수 있다.

- 방지: 영향을 방지하기 위해 프로젝트, 제품 또는 서비스를 변경한다.
- 감소: 영향을 최소화하기 위한 조치를 구현한다.
- 회복: 영향 이전의 상태로 복원하거나 복구하기 위한 조치를 취한다.
- 보상: 다른 완화 방법이 불가능하거나 효과적이지 않은 경우 현물 또는 기타 수단으로 보상한다.

즉, 방지를 우선시하는 접근 방식을 먼저 취하고 이것이 가능하지 않은 경우 영향을 줄이거나 완화하는 방법을 고려하는 것이다. 그러나 인권영향평가에 위 접근 방식을 적용할 때에는 주의할 점이 있다. 첫째, 취해진 모든 조치는 그 자체로 국제 인권 기준 및 인권 기반 접근 방식과 양립할 수 있어야 한다. 둘째, 구제가 포함되어야 한다. 보상과 구제는 동의어가 아니며 보상이 기본 구제가 되어서는 안 된다는 점을 이해하고 설명하는 것이 포함된다. 셋째, 인권영향은 ‘상쇄’ 대상이 될 수 없으며 이는 예를 들어 영향을 상쇄시킬 수 있는 환경 영향과 비교가 되는 부분이다.

기관은 사업 생태계의 정부 행위자, 동료 및 기타 제3자 행위자(개발자, 위원회, 투자자, 사용자 등)와 관련된 영향을 해결하기 위해 영향력을 행사해야 한다. 인권영향은 다양한 사업 기능 및 업무와 관련되기 때문에 다양한 부서가 인권영향 관리에 어떻게 관여할 수 있는지 고려하는 것도 중요하다(예: 법무 담당, 정책 담당, 기술 담당, 조달 담

당, 개인정보보호 담당관 등).

인권에 대한 부정적인 영향이 식별되고 영향 관리 계획이 수립되면, 조치를 이행할 것인지, 어떻게 이행할 것인지 후속 조치를 시행하여 식별된 영향을 효과적으로 해결하는 것이 중요하다. 이때 지속적인 모니터링은 영향 완화 조치가 효과적인지 여부에 대한 정보를 제공하고 그렇지 않은 경우 필요한 조정을 수행할 수 있도록 하며, 예상치 못한 영향도 식별할 수 있도록 한다.

구제책에 대한 접근 또한 영향 관리의 핵심 요소인 만큼, 고충 처리 체계를 마련하여야 한다. 고충 처리 체계를 통하여 인권영향평가 및 그 모니터링과 관련된 고충을 처리하고 디지털활동의 인권영향을 계속 식별할 필요가 있다.

○ 5단계 : 보고, 모니터링 및 점검

인권영향평가 방법 및 결과 보고서를 공개하고 소통하는 일은 인권영향평가 절차를 투명하고 책임성 있게 만드는 데 있어 중요한 부분이다. 최종 평가 보고서는 식별된 영향, 이를 해결하기 위해 취한 조치 및 조치 효과에 대한 모니터링을 문서화한 것이다. 덴마크 인권영향평가는 보고서 목차를 ①소개 ②방법론 ③맥락 설명 ④조사결과 및 조치로 구성할 것을 제안하였다.

이러한 보고서를 작성하고 공개하여 권리주체, 의무주체 및 기타 관련 당사자가 이용할 수 있도록 함으로써 대화와 책임성을 촉진할 수 있다. 네덜란드 아동노동실사법 등에서는 회사의 인권실사 활동을 보고하도록 제도적으로 요구하고 있다. 다만 보고서 공개 시 정보의 기밀성 및 민감성 등을 고려할 필요가 있다.

인권영향평가는 그 효과성 평가 및 지속적인 개선 조치까지 포함한다. 영향평가가 효과적이었는지 평가할 때에는 먼저 인권영향평가 절차 그 자체를 점검하면서 초기 목표를 충족했는지 파악하고 판단한다. 그 다음에는 최종 보고서 발행 후 후속적으로 예상하지 못한 영향, 회사 정책 및 관행에 대한 상당한 변경, 관련 현지 상황의 중대한 변화 등을 검토한다. 정례적인 검토는 인권영향평가 후 발생할 수 있는 문제를 해결하는 데 도움이 될 것이다.

○ 공통 : 이해관계자 참여

이해관계자 참여는 인권영향평가의 모든 단계에서 공통적으로 요구되는 구성요소이며, 이때 이해관계자는 권리주체, 의무주체 및 기타 이해관계자를 아우르는 개념이다.

이해관계자 참여 방식은 여러 가지가 있는데, 일방향적으로 사업에 대한 정보를 제공하는 방식, 양방향적 정보 교환 및 대화 방식, 특정 조치 후 의견을 수렴하는 방식, 상호 합의를 도출하기 위한 양방향적 협상 등으로 나누어 볼 수 있다.

계획 및 범위 설정 단계에서는 평가 절차에 참여해야 하는 이해관계자를 식별한다. 데이터 수집 및 상황 맥락 분석 단계에서는 권리주체와 그 대리인, 의무주체 및 기타 관련 당사자와 인터뷰 등을 통해 데이터를 수집한다. 이때 권리주체의 관점은 영향 분석 단계에서 영향의 심각도를 평가하는 데 필수적인 요소가 된다. 영향 예방, 완화 및 구제 단계에서 이해관계자는 부정적인 영향을 효과적으로 방지, 완화 및 구제하는 조치를 설계 및 시행하고 모니터링하는 데 의미 있게 참여해야 한다. 마지막으로 평가 결과는 이해관계자, 특히 권리주체에게 유의미하고 접근 가능한 방식으로 알려야 하며, 이들이 이후에 평가 절차에 참여하도록 해야 한다.

권리주체는 평가에 참여함으로써 관련 정보에 접근하고 디지털활동 및 그로 인한 영향을 더 잘 이해할 수 있을 뿐 아니라 자신의 인권과 이를 보호 및 존중해야 하는 의무주체의 의무와 책임에 대해 각각 알 수 있다.

의무주체 및 기타 관련 당사자의 참여 역시 포괄적인 평가를 보장하고 책무성을 향상시킨다는 점에서 필수적이다.

나. 평가의 기준

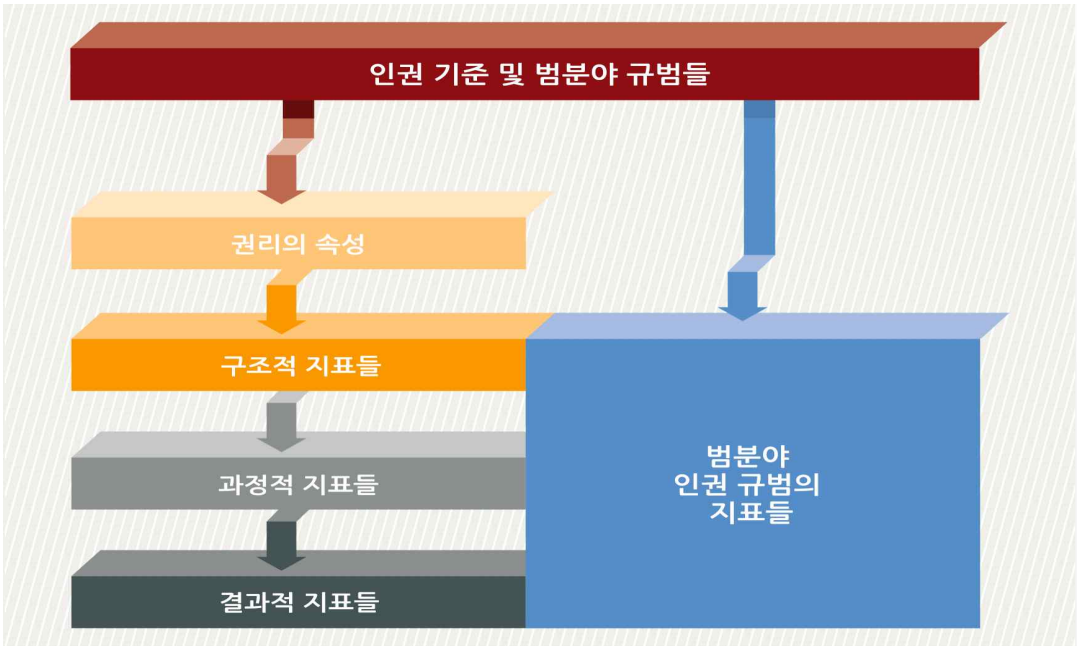
덴마크 인권영향평가의 특징은 평가의 내용 뿐 아니라 그 절차에 대해서도 기준을 두고 있다는 점이다. 관련하여 덴마크 국가인권기구는 인권영향평가의 10개 핵심 요소를 다음과 같이 정리하였으며, 아래 요소들이 인권영향평가의 고유성이라고 강조한다.

[표 7] 인권영향평가 절차 및 내용의 10개 핵심 요소

구분	핵심 요소	설명
절차	① 참여	영향을 받았거나 잠재적으로 영향을 받을 권리주체의 유의미한 참여가 영향평가 절차의 모든 단계(데이터 수집 및 맥락 분석, 영향 분석, 영향 예방, 완화 및 개선, 보고 및 평가 등)에 반영된다.
	② 차별 금지	참여 및 협의 절차에서 포용적이고 성인지적이며 취약하고 소외될 위험이 있는 개인 및 집단의 요구를 고려한다.
	③ 역량 강화	취약하고 소외될 위험이 있는 개인 및 집단의 유의미한 참여를 보장하기 위해 이들에 대한 역량 강화를 수행한다.
	④ 투명성	영향평가 절차는 영향을 받았거나 잠재적으로 영향을 받을 수 있는 권리주체를 적절히 참여시키기 위해 가능한한 투명해야 하며, 권리주체나 기타 참여자(인권단체 및 인권 활동가 등)의 안전과 안녕에 위험을 초래하지 않는다. 영향평가 결과를 적절히 공개한다.
	⑤ 책무성	영향평가팀은 인권 전문가의 지원을 받으며 영향평가, 예방, 완화, 관리에 대한 역할과 책임을 부여하고 적절한 자원을 제공한다. 영향평가는 권리주체의 자격과 관련 의무주체의 의무와 책임(예: 개발자, 디지털 제품 및 서비스를 구입하는 기업, 이를 사용하거나 적용하는 기업 또는 정부 기관)을 파악한다.
내용	⑥ 기준	인권 기준으로 영향평가의 기준을 구성한다. 영향 분석, 영향 심각도 평가 및 완화 조치 설계는 국제 인권 기준 및 원칙에 따라 수행된다.
	⑦ 영향 범위	평가는 사업이 야기하거나 기여한 실제적 및 잠재적 영향을 식별한다. 평가는 또한 기업의 운영, 제품 및 서비스 또는 사업 관계(계약 또는 비계약)를 통해 사업에서 직접적으로 관련된 영향을 고려한다. 평가는 누적적 영향 및 기존 문제를 분석한다.
	⑧ 영향의 심각도 평가	영향은 인권에 미치는 결과의 심각도에 따라 다루어진다. 여기에는 특정 영향의 범위, 규모 및 회복불가능성에 대한 검토가 포함되며, 권리주체 및 그 정당한 대리인의 견해를 고려한다.
	⑨ 영향 완화 조치	모든 인권영향이 다루어져야 한다. 영향을 다루기 위한 조치의 우선 순위를 정해야 하는 경우, 인권영향의 심각도가 핵심 기준이다. 식별된 영향을 해결할 때는 '방지-감소-회복-보상'의 완화 계층 구조를 따른다.
	⑩ 구제수단 접근	영향을 받는 권리주체가 디지털활동, 제품 또는 서비스는 물론, 영향평가 절차 및 그 결과에 대한 진정을 제기할 수 있는 방법이 있어야 한다. 기업이 영향평가 및 관리에서 영향을 받는 권리주체를 위한 구제수단 접근 권한을 보장하거나 협력한다.

다만 덴마크 인권영향평가는 시간 경과에 따라 인권 이행, 영향 및 변화를 측정하기 위한 지표의 개발이 여전히 새로운 분야라고 지적한다. 참고로 유엔 인권최고대표실이 개발한 인권 지표 프레임워크[88]의 경우 2단계 접근 방식을 취하였다. 첫 단계는 국제인권조약 등에서 규정한 규범적 내용(즉, 권리의 속성)을 설정하는 것이다. 예를 들어, 유엔 의사 및 표현의 자유 특별보고관은 "인공지능을 개발하고 도입하는 모든 민간 회사는 시민사회가 이에 대해 논평할 수 있도록 기회를 주어야" 하고, "기업은 엔지니어, 개발자, 데이터 기술자, 데이터 정제사, 프로그래머 및 기타 인공지능 수명 주기에 관여하는 이들에 대한 기업 정책 및 기술 지침에서 모든 사업 운영에는 인권적인 책임이 따른다는 사실을 반복적으로 안내해야" 하며, "플랫폼 서비스 약관은 보편적 인권 원칙에 기반해야" 한다고 말한 바 있다. 이러한 기준들이 인권 지표 개발에 사용될 수 있다. 두번째 단계는 이 규범적 내용에 기초하여 인권 이행을 측정하기 위한 지표를 구조적, 과정적 및 결과적 지표로 구분하여 설정하는 것이다.

[그림 12] 유엔 인권최고대표실 인권 지표 프레임워크



88) The Office of the United Nations High Commissioner for Human Rights(2012). Human Rights Indicators: A Guide to Measurement and Implementation, UN Doc. HR/PUB/12/5. <https://www.ohchr.org/Documents/Publications/Human_rights_indicators_en.pdf(접근일: 2022. 9. 10)>.

앞서 인권영향평가의 내용적인 요소 중 인권 기준 및 원칙을 디지털활동에 적용한 예시는 다음과 같다.

[표 8] 디지털활동에 대한 인권 기준 적용 예시

예시 시나리오	분석을 위한 인권 기준 및 원칙의 예시	영향을 받는 권리 및 관련 인권 규범
한 회사가 법치주의가 약하고 인권 활동가들이 박해받는 나라에서 수천 대의 카메라와 생체인식정보를 포함한 광범위한 데이터셋과 함께 사용될 얼굴 인식 기술을 제공하고 있다.	데이터 수집에 대한 이용자 동의가 없는 경우 이 사례에서 사생활에 대한 권리는 얼굴 기술 개발에 필요한 초기 데이터와 관련될뿐 아니라 이 기술이 적용될 때 수집 및 처리되는 데이터와 관련해서도 명확한 영향을 받는다. 그러나 다른 인권 기준도 영향을 받을 수 있다. 예를 들어, 이 시나리오는 결사의 자유와 표현의 자유에 영향을 미칠 수 있다. 개인들이 특정 단체에 가입하거나 정치 집회에 참여하기를 원하지 않게 될 수 있다. 얼굴 인식 기술에서 수집된 데이터가 개인을 체포 및 구금하는 데 사용되는 경우, 이 시나리오는 적법절차 및 공정한 재판의 권리에도 영향을 미칠 수 있다. 얼굴 인식 기술의 특정 적용은 또한 편향된 데이터에 기반하여 편향된 방식으로 적용될 수 있으며, 이는 평등권 및 차별받지 않을 권리에 부정적인 영향을 미칠 수 있다.	• 사생활의 권리 (세계인권선언 제12조, 자유권 규약 제17조) • 결사의 자유 및 노동조합 결성 및 참여의 권리 (세계인권선언 제20조, 자유권규약 제22조, 사회권규약 제8조) • 표현의 자유 및 정보의 자유 (세계인권선언 제19조, 자유권규약 제19조) • 공정한 재판의 권리 (세계인권선언 제11조, 자유권규약 제14조 및 제15조) • 차별받지 않을 권리 (세계인권선언 제7조 및 제23조, 자유권규약 제26조, 사회권규약 제2조)
한 기술 회사가 법 집행 기관 및 해당 형사 사법 제도를 위한 인공지능 도구를 개발했으며, 그 결과 많은 소수 종교 및 인종이 기소되었고 전보다 오랜 기간	차별받지 않을 권리는 국제인권법의 초석이다. 여기에는 직접적인 차별(‘공식적인’ 평등을 증진하기 위해 부당한 차별 대우를 다루는 문제)과 간접적인 차별(‘실질적인’ 평등을 증진하기 위해 외형적으로 중립적이어도 특정 보호 집단에 불이익한 조건을 다루는 문제)이 포함된다. 예를 들어, 유럽인권협약 차별금지 조항(제14조)은 특정 소수	• 차별받지 않을 권리 (세계인권선언 제7조 및 제23조, 자유권규약 제26조, 사회권규약 제2조) • 신체의 자유와 안전의 권리 [및 자의적인 체포로부터의 자유] (세계인권선언 제3조,

처벌받는 형을 선고받았다. 인공지능 도구를 사용한 결정에 대해 피고는 이의를 제기할 수 없다.	집단에 부정적인 영향을 미치는 중립적 조치를 금지하는 것으로 이해되고 있다. 무엇보다도, 이 사례의 회사는 인공지능 도구를 개발하는 데 사용되는 데이터가 편향되거나 차별적이지 않도록 보장해야 한다. 또한 인공지능 도구 자체를 차별적으로 적용할 가능성과 법 집행 기관과 판사가 차별적이지 않은 방식으로 인공지능 도구를 사용할 수 있는 능력을 고려하는 것이 중요할 수 있다. 또한 차별이 초기 관심사일 수 있지만, 일련의 인권이 인공지능 도구의 차별적 개발 및 적용과 그만큼 관련이 있을 것이다. 예를 들어 자의적 체포로부터의 자유 또는 법 앞의 평등에 부정적인 영향을 미칠 수 있다. 개인이 잘못된 형을 선고받을 경우, 이는 건강권, 교육권, 가족생활의 권리 등에 광범위한 영향을 미칠 수 있다.	자유권규약 제9조) • 건강권 (세계인권선언 제25조, 사회권규약 제12조) • 교육권 (세계인권선언 제26조, 사회권규약 제13조) • 가족생활의 권리 (세계인권선언 제16조, 자유권규약 제23조)
한 소셜 미디어 회사가 청소년과 아동에게 중독적인 것으로 보고된 제품과 서비스를 개발했다.	건강은 다른 인권의 행사에 불가결한 기본적인 인권이며, 모든 개인은 도달 가능한 최고 수준의 건강을 향유할 자격이 있다. 여기에는 신체적 건강과 정신적 건강이 모두 포함된다. 또한 아동은 특별한 보호가 필요한 취약 집단이고 그 건강한 발달을 촉진해야 한다. 소셜 미디어 중독은 아동의 복지와 정신 건강에 대한 권리에 심각한 영향을 미칠 가능성이 있다. 이 사례에서 회사는 제품 및 서비스가 중독 및 이에 상응하는 건강상 영향을 일으키지 않도록 재설계하는 조치를 취해야 한다. 나아가, 정신 건강에 대한 권리에 미치는 영향은 (초기 영향으로 인해 아동이 교육을 계속 받지 못한다면) 아동의 교육권을 비롯하여 아동의 권리와 관련하여 일련의 영향을 미칠 수 있다.	• 건강권 (세계인권선언 제25조, 사회권규약 제12조) • 아동의 건강권 (아동권리협약 제24조) • 교육권 (세계인권선언 제26조, 사회권규약 제13조) • 아동의 교육권 (아동권리협약 제28조)

덴마크 인권영향평가는 영향 분석의 기준으로는 ①영향의 유형 ②영향의 심각도 ③부정적인 영향을 제시한다.

먼저 영향의 유형과 관련하여 유엔 기업과 인권 이행지침은 잠재적이고 실제적인 인권영향을 검토할 것을 요구하는데, 이때 그 사업이 야기하는 영향, 사업이 기여하는 영향, 그리고 계약적 및 비계약적 관계를 비롯한 사업 관계를 통해 회사의 운영, 제품 또는 서비스와 직접 관련된 영향을 검토해야 한다. 나아가 평가는 사업이 같은 사람에게 연속적, 지속적, 결합적으로 영향을 미치는 누적적 영향 역시 분석하여야 한다.

[표 9] 디지털활동의 다양한 인권영향 유형

영향 유형	예시	잠재적인 영향을 받는 권리
회사의 고유 업무에서 '야기' 됨 (작위 또는 부작위)	• 한 부동산 회사가 채용 절차에서 지원자의 회사 내 성공 예상 순위를 매기기 위해 알고리즘을 구매하고 배치한다. 성공 예상 지수는 회사에서 제공한 과거 데이터를 기반으로 하며, 알고리즘은 지원자 분류에 보호대상 특성을 사용하지 않도록 개발되었다. 그러나 결과적으로 알고리즘은 다수 인종 남성 지원자만을 추천하였다.	평등권 및 차별받지 않을 권리
	• '스마트' 음성 비서를 개발하는 한 소프트웨어 개발사는 제품을 개선하기 위해 전체 이용자에 기반한 데이터를 대량으로 수집하고 있다. 이는 적절한 개인정보보호 및 충분한 설명에 입각한 동의 없이 수행되었다. 이용자는 공개되지 않은 제3자와 개인정보를 공유하는 데 동의해야만 이 제품을 사용할 수 있는데, 이는 설명에 입각한 동의가 부족하였음을 시사한다.	사생활의 권리
	• 정부와 계약한 한 데이터 엔지니어링 회사가 코로나19 감염병 유행시 사용할 접촉자 추적 앱을 개발했지만 수집 및 처리되는 민감정보를 보호하는 데 중요한 데이터 암호화 기능의 필요성을 검토하지 않았다.	사생활의 권리
	• 노조결성을 광범위하게 제한하는 국가에서 운영되는 한 은행이 노동자 만족도 및 유지율 향상에 도움이 될 신규 '스마트' 인적자원관리 시스템을 구매하고 적용한다. 신규 시스템은 자연어 처리를 사용하여 직원	결사의 자유

	간 내부 이메일을 분석하고 노동자의 정서상태를 평가한다. 노동자는 근로계약의 일부로 회사가 전자 메일에 접근하는 데 동의한다. 시스템 도입 후 회사 내 노조 설립률이 현저히 떨어졌는데, 이는 노동자들이 회사가 노조 활동을 더 쉽게 파악하고 노동자에 부정적인 영향을 미칠 수 있다고 우려하였기 때문이다.	
	• 한 통신사가 어느 나라 정부로부터 집권 정부에 대한 반대가 많은 특정 지역 인터넷을 폐쇄해 달라는 요청을 받는다. 여기서는 정부에 대한 평화적 시위가 소셜 미디어 플랫폼으로 조직되어 왔다. 통신사는 이런 시나리오가 발생하는 상황에 대한 정책이나 절차가 마련되어 있지 않아 아무런 문제제기 없이 정부의 요청에 충실히 응하고 있다.	표현의 자유, 평화적 집회 및 결사의 자유
누적적 영향을 비롯하여 회사의 고유 업무 또는 제3자를 통해 '기여' 함 (작위 또는 부작위)	• 한 소프트웨어 개발사가 상업적 목적으로 공용 소셜 미디어 플랫폼에서 '데이터를 스크랩' 할 수 있는 디지털 제품을 개발한다. 이 제품은 제3자에게 판매되고 제3자는 데이터를 스크랩하여 인터넷 이용자에 대한 정보를 정부에 제공하기 위해 이 제품을 사용한다. 정부는 이 데이터를 사용하여 정치적 반대자를 감시한다. 소프트웨어 개발사는 잠재적인 사용 사례와 관련 위험을 알고 있어야 한다.	사생활의 권리, 안전의 권리
	• 한 인공지능 개발사는 '효율적인 채용' 을 위해 자동화된 의사결정 알고리즘을 개발하고 이를 기업 고객에게 판매한다. 개발사는 차별적인 결과와 사생활의 권리에 대한 영향이 있을 수 있다는 점을 알고 있음에도 불구하고, 잠재적인 인권 위험과 이러한 위험을 방지할 수 있는 방법을 제품 구매자에게 알리지 않는다.	사생활의 권리, 평등권 및 차별받지 않을 권리
	• 널리 사용되는 소셜 미디어 플랫폼을 제공하는 한 기술 회사가, 플랫폼에 게시된 콘텐츠로 인종 폭력 및 분쟁이 촉발될 위험이 예상되는 국가에서 플랫폼에 게시된 콘텐츠를 선거 전에 검토하거나 조정하지 않는다. 회사는 플랫폼에 퍼지고 있는 콘텐츠 유형에 대해 알려 하지 않고 '방관하는' 접근 방식을 취한다.	건강권, 안전의 권리, 평등권 및 차별받지 않을 권리
	• 한 스타트업 소프트웨어 개발사가 슈퍼마켓 체인과 파트너십을 체결하고, 매장 내 제품 배치를 최적화하여 판매 촉진을 지원한다. 이를 위해 슈퍼마켓은 여러 매장에 카메라를 설치한다. 스타트업이 개발한 얼굴 특성화 기술은 고객의 감정 상태를 파악할 수 있다. 이 기술은 또한 단골의 쇼핑 행태를 식별하기 위해	사생활의 권리, 평등권 및 차별받지 않을 권리

	슈퍼마켓 체인의 단골 고객과 얼굴 특성화 기술을 연결하였다. 카메라의 설치 사실과 고객이 단골 프로그램에 등록하면 정보가 공유된다는 공지가 매장에 게시되어 있지만 매장 고객 중 소수만이 적용된 기술에 대해 알고 있다.	
	한 광고 기술(adtech) 회사가 다양한 소셜 미디어 플랫폼과 퍼블리셔로부터 이용자 데이터를 대량으로 수집하고 성적 취향을 비롯한 이용자 관심사를 기반으로 맞춤형 이용자 프로필을 개발했다. 회사는 광고주가 특정 '가치 청중'을 맞춤할 수 있는 기능을 홍보하고 광고 공간을 제3자가 쉽게 구매할 수 있도록 지원하였다. 이 광고 공간은 '동성애에 관심'이 있는 사람들을 혐오 표현의 표적으로 삼는 데 사용되었다. 온라인 발언이 오프라인 위협과 폭력으로도 이어진다.	평등권 및 차별받지 않을 권리, 사생활의 권리; 안전의 권리, 건강권, 생명권
계약 또는 비계약 사업 관계를 통해 그 운영, 제품 및 서비스와 '직접적으로 관련'됨	• 한 센서 회사가 자동차 회사에 다양한 센서를 판매했으며, 이 회사는 센서를 설치한 후 '스마트 차량공유' 애플리케이션으로 신규 시장에 진출하기로 결정했다. 요금 청구를 위해서는 탑승시 여러 데이터 포인트를 수집해야 하며, 이용자는 차량공유 프로그램에 가입할 때 이에 동의하였다. 효율성을 최적화하고 이용자 행태에 대해 자세히 알아보기 위해 운영자는 모든 탑승 시간 데이터를 수집하고 처리하기로 결정한다. 자동차 회사는 수집된 데이터의 전체 범위를 분석함으로써 운전자의 행동을 분석하고 기록할 수 있게 된다. 정부는 회사에 데이터 제공을 요구한다. 운전자 행태 정보 덕분에 정부는 야당 회의에 참가한 개인을 식별하고 탄압할 수 있다.	결사의 자유, 사생활의 권리
	• 한 인공지능 개발사가 신체 언어 인식 제품을 개발하고 법 집행 기관이 이 '기성품'을 구매한다. 법 집행 기관은 이 제품을 범죄 용의자 식별 용도로 적용한다. 이 제품의 사용으로 소수 민족 및 인종에 대한 부당한 체포가 증가한다.	평등권 및 차별받지 않을 권리, 이동의 자유, 신체의 자유와 안전의 권리
	• 한 사모펀드가 개인정보 보호법이 없는 국가에서 운영되는 바이오테크 회사에 투자한다. 인권실사 후 펀드는 회사의 개인정보보호 관행을 개선할 것을 권고한다. 그러나 외부 감사 결과 회사가 이용자 모르게 과도한 데이터를 수집 보유한 것으로 나타난다.	사생활의 권리

다음으로 식별된 영향에 대하여 조치의 우선순위를 결정하기 위해서는 영향의 '심각도'를 정의할 필요가 있다. 인권영향의 심각도 평가는 유엔 기업과 인권 이행지침에서부터 다음을 고려하여 왔다. 첫째, 회사가 야기 또는 기여하거나 직접적으로 관련된 인권의 모든 부정적 영향이 해결되어야 한다. 둘째, 모든 영향을 동시에 해결하는 것이 가능하지 않을 때 그 심각도 순서에 따라 해결되어야 한다. 가장 심각한 영향이 가장 먼저 해결되어야 한다. 셋째, 심각도는 영향의 '범위(scope)', '규모(scale)', '회복불가능성(irremediability)'에 의해 결정된다. 이때 범위는 영향이 얼마나 넓게 미치는지 또는 영향을 받는 사람 수에 대한 것이고, 규모는 영향의 중대성에 대한 것이고, 회복불가능성은 영향을 받은 개인을 영향 이전의 상황과 같거나 동등한 상황으로 복원하는 능력에 대한 것이다. 넷째, 통상 영향의 규모나 범위가 클수록 회복가능성이 높지 않다. 다섯째, 심각도 평가는 여성, 아동, 소수 인종, 장애인, 성소수자 개인 등을 비롯하여 취약하거나 소외될 위험이 높은 집단이나 개인별로 인권영향이 어떻게 다르게 미치는지에 특별한 주의를 기울여야 한다.

잠재적인 인권영향의 심각도가 여럿 식별되었고 이를 동시에 처리할 수 없는 경우, 조치 우선순위 선정을 위해 두 번째로 고려할 사항은 잠재적 영향이 실현될 가능성이다. 유엔 인권최고대표의 <B-테크 기초 자료>에 따르면 디지털활동에서는 다음을 고려하여야 한다.

- 개발자 또는 사용자의 관심사, 동기 및 인센티브: 위험을 초래할 수 있는 방식으로 제품이나 서비스를 사용하거나 오남용하는 것이 사용자에게 이익이 되는가? 위험을 초래할 수 있는 방법으로 제품이나 서비스를 개발하는 것이 개발자의 이익에 부합하는가?
- 사용자의 기술 노하우 및 기능: 사용자의 노하우(또는 노하우 부족)가 식별된 사용 사례와 부정적인 영향이 발생할 가능성에 변화를 가져오는가? 부정적인 사용 사례의 실제 실현을 불가능하게 만드는 기술 장벽이 있는가?
- 설계에서 인권을 고려하는 개발자의 능력: 개발자가 인권 개념을 이해하고 그 범위 내에서 식별된 잠재적 인권영향을 방지할 수 있는 방식으로 디지털 제품이나 서비스를 설계할 수 있는가?
- 적용 정책 및 법률: 정부 정책과 법률이 사용 사례를 실제로 발생시킬 가능성을 어느 정도 높이는가?

덴마크 인권영향평가는 영향 심각도 평가와 관련하여 추가로 주의해야 할 5가지 사항을 다음과 같이 꼽았다.

첫째, 영향의 심각도를 정할 때 영향을 받거나 받을 수 있는 개별 권리주체 및 지역사회 구성원, 또는 이들을 대표하는 단체와 대화하여야 한다.

둘째, 영향의 심각도를 정할 때 취약성을 필수적인 요소로 고려하여야 한다. 예를 들어, 어떤 기업이 노동자의 정서상태를 분석하기 위해 자연어 처리를 사용하는 "스마트한" 인적자원 도구를 사용하는 것에 대해 직원 동의를 구하는 경우, 노조에 가입되어 있지 않은 시간제 노동자가 노조에 가입되어 있는 정규직 노동자에 비해 동의를 거부할 수 있는 권한이 적기 때문에 시간제 근로자에게 미치는 영향이 더 클 수 있다. 또다른 기업이 신용평가에 자동화된 의사결정을 사용하는 경우, 신용점수에 기초하여 이루어지는 차별적이거나 편향적인 결정은 부유한 개인보다 가난한 개인에게 더 심각할 수 있으며, 결정 사항에 대해 이의를 제기하기 위해 동원할 수 있는 자원 면에서도 마찬가지일 것이다.

셋째, 범위(영향 대상자 수)를 검토할 때 영향을 받는 개인의 절대적인 수치뿐 아니라 영향을 받는 개인이 누구인지 자세히 고려하는 것이 필수적이다. 예를 들어, 어느 디지털 제품의 실제적 영향을 받는 사람들의 절대수는 100명 중 다섯 명에 불과하지만, 그 다섯 명이 항상 여성단체 활동가인 경우, 이는 특정 집단의 사람들에 대한 체계적 탄압일 수 있으므로 분석에 나타나야 한다.

넷째, 평가 과정이 적절한 정보 하에 진행되기 위해서는 인권 전문지식(인권전문기구, 인권 전문가 등)의 역할이 중요하다.

다섯째, 심각도는 절대적인 개념이 아니며, 심각도에 대한 보편적인 임계값도 없다. 그보다 심각도 평가는 영향 식별과 밀접한 관련이 있으며, 전문적인 판단, 이해관계자 대화, 영향의 상관관계(예를 들어 통상 사생활권리의 영향은 안전의 권리 및 집회시위의 권리 등 다른 권리에 대한 '관문'이 된다)는 물론 장기적인 결과에 대한 분석을 통해 이루어진다.

이러한 검토를 거쳐 덴마크 인권영향평가는 디지털활동 인권영향의 심각도 평가 지표를 다음 표와 같이 제시하였다.

[표 10] 덴마크 인권영향평가 심각도 평가 지표

항목	지표	심각도	비고
범위	영향 영역 전체 인구의 20% 이상 또는 식별된 집단의 50% 이상	A	인권 관점은 특정한 개인들이 향유하고 행사하는 인권과 자유를 강조한다. 따라서 범위(영향을 받는 사람의 수)를 고려할 때 절대적인 숫자만이 아니라 영향을 받는 개별 이용자 및 기타 권리주체가 누구인지 보다 정확하게 검토한다. 일부 영향은 수치적으로는 작을 수 있지만 비례적으로 더 큰 타격을 받는 특정 권리주체 집단에 편향될 수 있다. 예를 들어, 디지털 커뮤니케이션 플랫폼의 이용자 중 0.1%만이 영향을 받을 수 있지만 이것이 소수 종교인의 25%라면 후자가 전자보다 관련성이 더 높다. 집단 식별은 잠재적으로 영향을 받는 사람들(여성 이용자 또는 남성 이용자 등)을 세분화하는 상황별 방법론이 될 수 있다.
	영향 영역 전체 인구의 10% 이상 또는 식별된 집단의 10-50%	B	
	영향 영역 전체 인구의 5% 이상 또는 식별된 집단의 10% 미만	C	
규모 (취약성의 고려 등)	사망 또는 정신적·육체적 건강에 악영향을 미쳐 삶의 질 및 수명을 현저히 감소시킬 수 있다. 여기에는 안전의 권리 또는 건강권과 관련한 심각한 영향으로 이어지는 사생활의 권리에 대한 영향이 포함된다.	A	영향의 규모 또는 심각성을 고려하는 데 취약성이 필수적인 사항이어야 한다. 이는 변화에 대응하는 능력을 비롯하여 개인의 특정 상황이 해당 개인에게 얼마나 '심각한' 영향을 미칠 수 있는지에 영향을 미칠 수 있기 때문이다. 평가자는 취약성을 규모의 일부로 고려할 뿐 아니라 분석에서 취약성이 고려된 방식을 명확하게 보여주기 위해 취약성을 별도의 지표로 항목화할 수 있다.
	기초생활필수품(교육, 생계 등)에 대한 명백한 인권침해. 영향평가 절차에서 식별된 집단 또는 주제별 전문가가 높게 평가한 것으로 식별된 문화적, 경제적, 자연적, 사회적 측면에 대한 영향 영향평가 과정에서 생계, 건강 또는 안전에 우선순위가 높은 것으로	B	

	식별된 공공 서비스 전달과 관련된 부정적 영향. 예를 들어 공공의료 서비스에 접근하는 개인을 지원하는 인공지능 챗봇이 노인을 돕는 효율성이 떨어져 일반인보다 적절한 건강 조언을 덜 받는 경우		
	그밖의 영향 모두	C	
회복 불가능성	어려움: 영향의 특성상 구제하기 어렵거나 불가능하다. 복잡한 기술 요구사항으로 인해 개선이 어렵다. 식별된 집단에서 구제책 수용을 거의 발견할 수 없다. 영향에 관련된 사업 파트너가 영향을 구제할 수 있는 능력이 낮다. 영향으로 인한 손실을 대체할 수 있는 실행 가능한 대안이 없다.	A	디지털 감시 기술로 인해 체포되어 고문을 받고 개인의 건강권이 영향을 받은 경우 본질적으로 구제할 방법이 없다. 비즈니스 모델이 광범위한 데이터 수집 및 공유에 의존하는 경우 공유된 데이터와 관련된 영향을 회복시키는 것은 한 회사의 통제권 밖에 있기 때문에 구제되기 어려울 것이다.
	보통: 영향의 특성상 구제가 가능하지만 쉽지 않다. 영향을 구제하기 위한 기술적 요구사항이 보다 간단하다. 영향을 받는 식별된 집단이 구제책을 수용한다. 일부 기능 개발이 지원되는 경우 영향에 관련된 사업 파트너가 구제수단을 제공할 수 있다.	B	'스마트 채용 시스템'의 사용자가 인권에 부정적인 영향을 미치지 않는 방식으로 시스템을 사용할 수 있으며 시스템 사용에 대한 적절한 동의를 구하는 방법에 대해 교육훈련을 받는다.
	쉬움: 영향의 특성상 구제하기 쉽다. 영향을 구제하기 위한 기술적 요구사항이 간단하다. 영향을 받는 식별된 집단이 구제책을 수용한다. 사업 파트너가 영향을 구제할 수 있는 능력을 가지고 있다.	C	특정 데이터 브로커와 협력하는 위험을 확인한 통신 회사가 개인정보보호 문제로 인해 해당 브로커와 협력하지 않기로 결정할 수 있다.

위 심각도 지표를 가상의 시나리오에 적용해본 예시는 다음과 같다. 회사가 어떤 국가의 사법부가 도주 및 재범 위험을 계산하여 판결을 지원하는 데 사용하는 알고리즘을 개발하였고, 그 대상은 자동화된 위험평가에 대해 이의를 제기할 수 없다고 할 때, 이 알고리즘은 권리주체의 적법 절차와 공정한 재판을 받을 권리에 잠재적인 영향을 미치며, 범주별 심각도는 다음 표와 같다.

[표 11] 덴마크 인권영향평가 심각도 평가의 예시

항목	심각도
범위	B: 판결만을 보조하는 고도로 정확한 알고리즘의 경우 영향을 받는 사람의 절대적인 수는 적을 수 있지만, 사례를 자세히 살펴본 결과 주로 영향을 받는 사람들이 선주민으로 나타났고 취약 집단으로 식별되었다.
규모	A: 공정한 재판을 받을 권리에 대한 영향으로 상당한 규모의 인권영향이 있다. 알고리즘의 지원으로 이루어진 판결을 받은 모든 사람에게 적용된다. 그러나 영향을 받는 선주민과 관련한 영향은 훨씬 더 크며, 이는 그 영향이 본질적으로 차별적이기 때문이다.
회복불가능성	B: 판결의 성격에 따라 상황을 완전히 구제하지 못할 수도 있다. 예를 들어, 미래의 직업 기회가 제한되어 평생 영향을 미칠 수 있다. 형은 오랫동안 정신 건강에 영향을 미칠 수 있으며, 수년간 가족생활을 영위할 수 없으며, 반드시 회복될 수 있는 것이 아니다. 그러나 영향의 일부는 차별적 판결을 변경하여 구제될 수 있다.
종합 평가	높은 심각도의 영향으로 간주될 수 있다. 이는 특히 취약 집단에 영향을 미치는 지속적인 영향이며 일부 사례(과도한 징역형)는 예를 들어 건강과 생계 및 가족 생활에 대한 권리와 관련한 영향으로 인해 구제될 수 없다.

마지막으로 인권영향평가는 부정적인 영향에 주목하여야 한다. 부정적인 인권영향에 관여한 많은 기업이 해당 디지털활동의 편익에 주의를 돌리고자 하는데, 유엔 기업과 인권 이행지침은 긍정적인 기여도로 부정적인 영향을 상쇄할 수 없으며 각각 검토하여야 한다고 지적한 바 있다. 덴마크 인권영향평가는 행정 효율성 향상, 이용자 경험 최적화, 정보접근성 향상 등 디지털활동의 긍정적인 기여와 그 부정적인 영향을 별개로 구분할 때, 인권영향평가가 주요하게 초점을 맞추고자 하는 부정적인 인권영향과 그에 대한 각 의무주체의 책무성 및 완화 조치를 제대로 살필 수 있다고 강조한다.

다. 다른 규제 메커니즘과의 관계

국가는 개인정보보호 영향평가 등 여러 영향평가 제도를 병행하여 운용하고 있다. 그 대표적인 예시가 유럽연합 GDPR에 따라 고위험 개인정보처리에 대하여 시행하는 개인

정보보호 영향평가이다.

덴마크 국가인권기구는 개인정보보호 영향평가와 인권영향평가 사이에는 유사한 부분이 있지만 상호 완전히 수렴하지는 않는다고 지적한다. 특히 인권영향평가가 개인정보보호 영향평가에 병합되거나 통합될 경우 인권영향평가의 요소가 손실될 위험이 있다고 우려한다. 개인정보보호 영향평가가 인권영향평가의 방법론을 반영할 경우 개선될 수 있는 부분도 있는데, 예를 들어 인권영향평가의 기준대로 평가 결과를 공개한다면 개인정보보호 영향평가의 정보 제공 및 투명성을 증진할 수 있을 것이다.

인권영향평가는 단독해서 실시하거나 타 영향평가와 결합하여 실시할 수 있지만, 어느 것이 인권 평가에 바람직한지의 문제는 상황별로 다를 수 있다. 결합형의 경우, 기존 영향 및 위험 관리 구조를 활용할 수 있고, 평가 피로를 방지할 수 있으며, 개인정보보호・사생활 보호・윤리적 원칙과 인권영향의 상호 관련성에 대한 분석을 촉진할 수 있을 뿐 아니라, 각 분야의 강점을 살릴 수 있다는 점에서 이점이 있다. 다른 한편 단독형의 경우, 인권 문제 회피를 방지할 수 있고, 인권 전문성을 광범위하게 활용할 수 있으며, 다양한 이해관계자의 학습 및 역량구축을 꾀할 수 있다는 이점이 있다. 덴마크 국가인권기구는 이 두 가지 접근법의 절충안으로 분야별 영향평가(Sector Wide Impact Assessment)를 제시한다. 디지털활동, 인공지능 등 특정 분야별로 인권영향평가를 포함하는 영향평가를 실시할 경우 상호 조정이 용이하고 동일한 맥락에서 유사한 인권영향을 식별하고 다루기에 효율적이라는 것이다.

라. 운영상 쟁점

1) 평가 수행 주체

덴마크 인권영향평가의 적용 대상은 공공기관과 민간의 개발 사업자 또는 구매 사업자이며 인권실사의 일부로서 인권영향평가를 자율적으로 실시하도록 하였다. 다만 인권영향평가의 실시를 담당하는 팀이 기관과 독립적일 때 예방 및 완화 조치의 관측과 권장사항 도출의 정당성을 보장하고 기관의 인권 담당 직원을 적절하게 지원할 수 있다. 인권영향평가팀이 전적으로 내부 직원으로 구성될 경우 평가의 독립성을 제한하게 되지만, 외부인으로는 기관 내부에 대한 지식과 전문성이 결여될 수밖에 없다. 따라서 평가

팀, 회사 임직원, 기타 이해관계자가 함께 평가를 위한 팀을 구성하는 것이 바람직하다.

2) 이해관계자 참여

덴마크 인권영향평가는 모든 단계에서 공통적으로 이해관계자 참여를 요구한다. 덴마크 국가인권기구는 인권영향평가에서 참여를 요구하는 이해관계자가 보편적으로 정해져 있지는 않지만, 권리주체, 의무주체 및 기타 이해관계자의 참여가 필수적이라고 강조한다. 인권영향평가가 이해관계자 참여 없이 탁상 연구에 그친다면 인권영향에 대한 철저한 평가가 가능하거나 효과적이지 않다는 것이다.

또한 유엔 기업과 인권 이행지침에서 지적한 대로 영향을 받았거나 잠재적으로 영향을 받을 사람으로서 권리주체의 유의미한 참여가 영향평가 절차의 모든 단계에 반영되어야 한다. 취약하고 소외될 위험이 있는 개인 및 집단의 경우 유의미한 참여를 보장하기 위해 이들에 대한 역량을 강화하고 관련 인권단체나 노동조합 등 적절한 대리인의 참여가 이루어질 필요가 있다. 예를 들어 '스마트 채용 시스템'이 여성에게 차별적 영향을 미칠 수 있음이 확인되면 여성 권리주체 및 여성단체가 데이터 수집 과정에 참여해야 한다.

[그림 13] 인권영향평가 이해관계자 권한 매핑

특히 덴마크 인권영향평가는 이해관계자를 식별하는 과정에서 [그림 13]과 같은 모형으로 ‘이해관계자 권한 매핑’을 실시해볼 것을 권장한다. 이해관계자가 디지털활동에 미치는 영향을 Y축으로 삼고, 디지털활동이 이해관계자에게 미치는 영향을 X축으로 삼아 관련된 이해관계자를 채워보면, 권한이 적고 영향은 가장 많이 받은 취약한 권리주체를 식별할 수 있다.

3) 평가 결과 공개

덴마크 인권영향평가는 영향을 받았거나 잠재적으로 영향을 받을 수 있는 권리주체를 적절히 참여시키기 위해 절차를 가능한한 투명하게 운영해야 하며, 영향평가 결과 또한 적절히 공개되어야 한다고 밝힌다.

공개된 인공지능 관련 인권영향평가 사례로는 2019년 구글이 실시한 얼굴인식 기술에 대한 평가를 들 수 있다. 공개된 평가 요약본에 따르면 구글은 미디어 및 엔터테인먼트(M&E) 산업 분야에서 사용되는 자사 얼굴인식 기술에 대한 인권영향평가를 의뢰하였으며, 자사 “유명인사 인식 API 개발”에 인권 정보를 제공하고자 하였다. 이 API는 구글이 M&E 산업 기업 고객으로 하여금 “Google이 라이센스를 부여하고 클라우드 인공지능 제품 포트폴리오의 일부로 사용할 수 있는 유명인 이미지 데이터베이스를 사용하여 프레임별 또는 장면별 수준에서 유명인사를 식별”할 수 있도록 한다. 평가에는 잠재적으로 영향을 받는 이해관계자가 참여한 것은 물론 독립적인 전문가의 자문을 받았다고 한다.

4. 네덜란드 기본권 알고리즘영향평가

네덜란드 세무 당국은 알고리즘 시스템에 기반하여 저소득 가정 부모와 돌봄 노동자 수만 명을 아동복지 급여 부정수급 혐의로 고발하였으나, 이 의사결정이 소수 인종에 편향적이었다는 사실이 2021년 알려지면서 큰 논란을 빚었다. 특히 이 시스템은 특정 개인을 부정수급 혐의자로 분류한 이유를 제공하지 않아 공공부문의 투명성, 책무성, 감독이

결여되었다는 비판을 받았다.[89]

2022년 4월 5일, 네덜란드 하원은 정부에 인권영향평가를 의무화할 것을 요구하는 결의안을 채택했다. 이 결의안은 네덜란드 정부가 개발한 인권영향평가를 의무화하면 알고리즘 남용을 예방할 수 있다고 지적하면서, 공공기관이 알고리즘을 사용하여 사람에 대한 평가 또는 결정을 내릴 때 사전 인권영향평가를 의무화하고 그 평가 결과 또한 가능한 한 공개하도록 의무화할 것을 요구하였다.[90]

관련하여 네덜란드 내무부는 Utrecht 대학에 위탁하여 알고리즘 인권영향평가도구인 기본권 알고리즘영향평가(Fundamental Rights and Algorithms Impact Assessment)[91]를 개발하였다. 이 평가 목표는 인권과 결과가 예측되지 않거나 인권에 부정적인 영향을 미치는 알고리즘의 도입을 방지하는 데 있다.

가. 평가의 절차

네덜란드 기본권 알고리즘영향평가는 인공지능 시스템을 개발하거나 도입하는 초기 단계에서 논의하고 해결하여야 할 인권 쟁점에 대한 질의로 구성되어 있으며 다음 단계를 거쳐 답하도록 하였다. 1부는 준비 단계로 알고리즘이 사용되는 이유와 그 효과가 무엇인지 판단한다(Why). 2부는 입력 및 처리 단계로 알고리즘 시스템의 개발에 관한 것이다(What). 즉, 이 단계에서 알고리즘을 개발하는 데 어떤 데이터가 사용되는지, 알고리즘이 어떤 형태여야 하는지를 결정한다. 특히 데이터에 대해서는 특정 유형의 데이터 및

89) Amnesty International(2021). Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms. <https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/(접근일: 2022. 8. 15)>.

90) MOTIE VAN DE LEDEN BOUCHALLIKH EN DEKKER-ABDULAZIZ(Nr. 835). Voorgesteld 29 maart 2022; Utrecht University(2022. 4. 8). Dutch House of Representatives endorses mandatory use of Human Rights and Algorithms Impact Assessment. <https://www.uu.nl/en/news/dutch-house-of-representatives-endorses-mandatory-use-of-human-rights-and-algorithms-impact(접근일: 2022. 8. 15.)>; European Center for Non-for-Profit Law(2022). Netherlands Sets Precedent for Human Rights Safeguards in Use of AI. <https://ecnl.org/news/netherlands-sets-precedent-human-rights-safeguards-use-ai(접근일: 2022. 8. 15.)>; Knowledge Centre Data & Society(2022). Netherlands - Fundamental Rights and Algorithms Impact Assessment(FRAIA). <https://data-en-maatschappij.ai/en/policy-monitor/nederland-impact-assessment-mensenrechten-en-algoritmes(접근일: 2022. 8. 15)>.

91) Government of the Netherlands(2022), 질의 문항 번역은 부록 IV 참조.

데이터 소스의 사용에 대하여 질의하고, 알고리즘에 대해서는 알고리즘의 작동 및 투명성에 관해 질의한다. 3부는 출력, 구현 및 감독 단계로, 알고리즘을 사용하는 방법에 관한 것이다(How). 즉, 알고리즘이 생성하는 결과물이 무엇인지, 그 결과물이 정책 또는 의사결정에서 어떤 역할을 할 수 있는지, 이를 어떻게 감독할 수 있는지 질의한다. 이때 모든 단계에서 기본권 침해 위험을 식별하고 그 정당성을 평가할 필요가 있다. 이에 기본권 알고리즘영향평가는 앞서 3단계 질의에서는 알고리즘 사용에 관한 의사결정 단계에 대하여 질의하고, 마지막으로 4부에서 기본권에 대하여 광범위하게 질의한다.

○ 1단계: 준비 단계 - 왜 하는가?

기본권 알고리즘영향평가 1부는 알고리즘 사용의 이유 및 목적에 대한 포괄적인 질의를 먼저 논의하도록 함으로써 이후 기본권에 대한 요구사항이나 기본권에 미치는 잠재적 영향에 대한 질의로 이어지도록 하였다.

[그림 14] 네덜란드 기본권 알고리즘영향평가 절차

우선 (1.1)항목에서 이유 및 문제 정의에 대하여 질의한 것은, 문제를 해결하는 데 다른(비디지털) 수단을 사용할 수 있음을 이해하고 알고리즘을 사용하는 것이 바람직하거나 필수적인 이유를 판단할 필요가 있기 때문이다. 또한 이 단계 논의에서 사업팀장, 해당 분야 전문가(직원) 등 내부 이해관계자 뿐 아니라 시민이나 이익단체가 참여할 수 있다면 다양한 관점의 조기 통찰력을 얻고 알고리즘 사용에 대한 지지를 구축할 수 있다고 강조한다. (1.2)항목은 의도된 효과, 즉 목적에 대하여 질의하면서 가능한 한 구체적으로 답하도록 하여 이후 단계에서 이를 측정하고자 하였다. 특히 개인정보 처리가 수반되는 경우 개인정보보호 영향평가 또한 목적 구속성을 중요한 원칙으로 삼고 있다는 점 또한 강조하였다. (1.3)항목에서 관련된 공공가치에 대하여 질의하는 이유는 이후 알고리즘 사용의 결과가 공익과 기본권에 균형적으로 영향을 미치는지 평가하기 위함이며, (1.4)항목에서 묻는 법적 근거는 이후 기본권이 제한될 것으로 예상되는 경우 중요한 평가 대상이다. (1.5)항목은 이해관계자 및 책임성에 대하여 질의하는데, 이때 의도치 않은 결과가 발생하였을 경우 적시에 조정하거나 완화 조치를 취할 수 있는 현재의 업무 분장 뿐 미래의 할당도 생각하도록 하였다. 만약 책임성을 충분히 보장하는 것이 불가능한 것으로 판명될 경우 알고리즘을 사용해서 안되기 때문에 출구전략이 중요하다.

○ 2단계: 입력 및 처리 단계 - 무엇을 하는가?

기본권 알고리즘영향평가 2부는 다시 데이터에 대한 하위 항목 2A부와 알고리즘에 대한 2B부로 나뉜다. (2A.1)항목은 데이터에 대한 질의 전에 알고리즘의 유형을 먼저 질의하면서 어떤 유형의 알고리즘을 사용할 것인지에 대한 구상이 없으면 이후 질의를 답하기 어렵다고 지적한다. 만약 어떤 알고리즘 유형을 사용할지 대략의 구상도 없다면 평가를 일단 중단하고 추후 다시 실시할 것을 권장한다.

데이터 원천 및 품질에 대한 (2A.2)항목과 데이터 편향성/가설에 대한 (2A.3)항목의 질의들은 데이터 품질이 알고리즘 결과물에 결정적이기 때문에 중요하다. 데이터는 완전하고 정확해야 하며 그 편향성이 알고리즘의 결과물에 미치는 영향을 수정, 극복, 완화하여야 한다. 또 데이터는 대상이나 알고리즘이 사용될 맥락을 적절하게 대표할 수 있어야 한다. 해당 알고리즘이 학습 데이터를 사용하는 경우 그 출처와 품질을 조사해야 한다.

보안 및 보관에 대한 (2A.4)항목은 개인정보의 익명화 또는 가명화 등으로 데이터의 식별을 방지하고 로그기록, 접근통제, 보관 규칙을 준수하는지 등 안전조치에 대한 사항을 묻는다.

알고리즘 처리와 관련한 2B부는 우선 (2B.1)항목에서 알고리즘 유형에 대하여 질의하면서 특히 비자기지도 학습 알고리즘(non-self-learning algorithm)과 자기지도 학습 알고리즘(self-learning algorithm)을 구분하였다. 이들 유형을 구분한 것은 알고리즘 사용에 대해 서로 다른 질의를 불러오기 때문이다. 이어서 소유권 및 통제권에 대한 (2B.2)항목은 알고리즘이 외부에서 개발된 경우, 알고리즘의 소유권 및 관리 권한에 대해 명확한 합의가 이루어졌는지, 그 합의 내용은 무엇인지 묻는다. 이에 대한 답변은 알고리즘의 설명가능성과 관련이 있으며, 알고리즘이 제3자에 의해 개발된 경우에도 실제로 알고리즘을 사용하는 기관에서 알고리즘이 어떻게 작동하는지 설명할 수 있어야 한다. (2B.3)항목은 알고리즘 정확성을 질의하면서 참양성, 위양성, 참음성, 위음성 등 정확도를 평가하고 해당 정확도가 알고리즘이 사용되는 상황에서 수용 가능한지 여부를 조사하고 논의하도록 하였다. (2B.4)항목은 투명성과 설명가능성에 대한 질의이다. 이 항목에 대한 해설에서는 ‘투명성’에 대하여 적용 중인 알고리즘의 방법론(의사결정 트리, 신경망), 소스코드, 알고리즘이 학습한 방법, 데이터, 입력 변수, 매개변수 및 사용된 임계값 등에 대한 이해에 관한 것이라고 소개하는 한편, ‘설명가능성’에 대하여는 데이터 분석 결과와 그 분석 결과가 어떻게 나왔는지 이해할 수 있는 언어로 설명할 수 있는 능력이라고 구분하여 소개하였다. 설명가능성과 투명성이 필요한 정도는 (1)결정, 결과 및 시민에 대해 미치는 알고리즘의 영향, (2)의사결정의 자율성 수준(인간의 참여가 보장되는 정도), (3)알고리즘의 유형 및 복잡성에 달라진다.

○ 3단계: 출력, 구현 및 감독 단계 - 어떻게 하는가?

기본권 알고리즘영향평가 3부는 알고리즘의 결과물에 대한 질의로서 먼저 (3.1)항목에서 알고리즘 결과물에 기반한 의사결정이 이루어질 경우 기대치를 충족할 수 있을지 그 명확한 상을 구하는 데 중점을 둔다. 이 질의에 답하려면 앞서 이유(1.1), 목적(1.2) 및 가치(1.3)에 대한 질의에 답한 내용을 고려해야 한다. 의사결정에서 인간의 역할을 묻는 (3.2)항목은 신뢰할 수 있는 인공지능을 위해서는 의사결정이 완전히 자동화된 경우에도

일정한 수준의 인간 개입이 필요함을 시사한다. 또한 책임있는 의사결정을 위하여 현재뿐 아니라 미래에도 직원의 전문성을 보장하고 적절하게 업무를 분장하는 한편, 제3자와 협력할 때 알고리즘에 대한 유지관리를 보장해야 한다. (3.3)항목은 알고리즘의 효과에 대하여 질의하면서 특히 1부에서 확인된 공공 가치와 개인의 이익에 비추어 알고리즘 결과물에 의한 영향의 크기와 특성을 판단하고자 한다. (3.4)항목은 의사결정 및 시민 참여 절차에 대하여 묻고 (3.5)항목은 맥락에 대하여 묻는다. 특히 (3.5)항목은 알고리즘이 의도되거나 학습한 것과 다른 맥락에서 사용되는 경우 부정확하거나 편향된 결과물을 낳을 수 있다고 지적한다. 맥락이 달라지면 알고리즘을 뒷받침하는 가설이 더 이상 적용될 수 없다. 반대로 알고리즘을 뒷받침하는 가설이 변경되면 과거 선택했던 맥락에 알고리즘을 더 이상 적용할 수 없다. (3.6)항목은 개방성에 대한 질의이다. 기관은 알고리즘의 작동 및 중요성에 대해 완전히 개방적이거나 완전히 폐쇄적일 수 있지만, 그 사이에도 많은 형태의 선택지가 있을 수 있다. (3.7)항목은 알고리즘의 작동 및 효과에 대하여 평가, 감사, 보장 체계로서 책무성을 보장하는 문제에 관한 것이다. 알고리즘의 작동은 계속 검증될 필요가 있는데, 이러한 검증 절차로는 내부 평가, 감사 및 보장 절차(알고리즘을 적용한 정부 기관에서 수립한 절차 등)는 물론 외부 절차(외부 감독기구의 감독 등)도 고려할 수 있다.

○ 4단계: 기본권 로드맵

기본권 알고리즘영향평가 4부는 <기본권 로드맵>이라는 제하로 사용할 알고리즘이 기본권에 영향을 미치는지 여부를 식별하고, 기본권 행사에 대한 간섭을 방지하거나 완화할 수 있는지, 기본권 간섭이 수용 가능한지 등을 검토하고자 한다. 이를 위해 7단계의 검토를 거치도록 하였는데 각 단계의 검토 주제는 다음과 같다.

네덜란드 기본권 알고리즘영향평가 7단계 기본권 검토
(7.1) 기본권: 알고리즘이 기본권에 영향을 미치는가? 또는 영향을 미칠 위험이 있는가?
(7.2) 구체적인 법률: 고려해야 할 기본권과 관련하여 구체적인 법률이 적용되는가?

(7.3) 심각도 정의: 이 기본권이 얼마나 심각하게 침해되고 있는가?
(7.4) 목적: 알고리즘을 사용하여 어떤 사회적, 정치적 또는 행정적 목적을 달성하려 하는가?

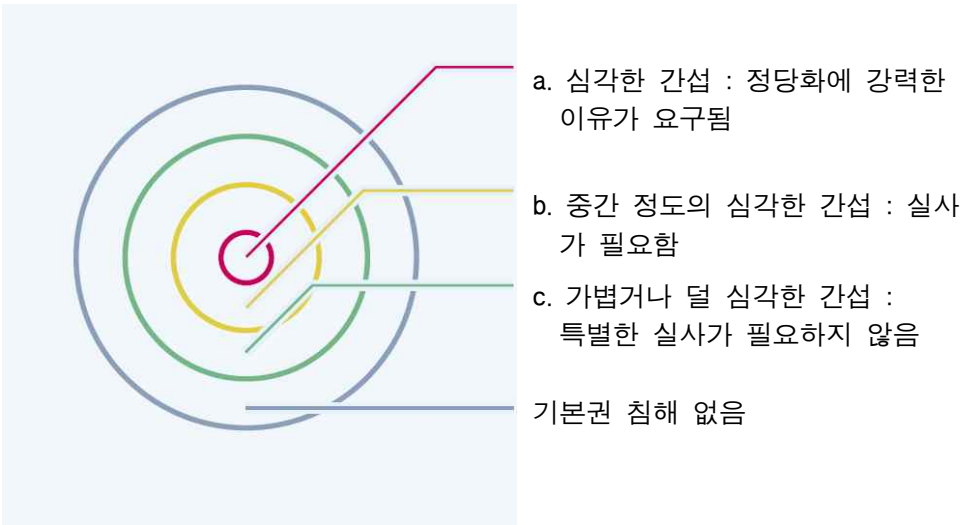
(7.5) 적합성: 특정 알고리즘을 사용하는 것이 이러한 목적을 달성하는 데 적합한 수단인가?

(7.6) 필요성 및 보충성: 이 목적을 달성하기 위해 특정 알고리즘을 사용하는 것이 필요한가? 이를 위해 사용할 수 있는 다른 수단 또는 완화 조치는 없는가?

(7.7) 균형성 및 비례성: 기본권에 영향을 미치는 것을 결국 정당화할 만큼 이 목적이 충분히 중요한가?

여기서 특히 주목할 부분은 (7.3) 단계에서 알고리즘이 기본권에 미치는 영향의 심각도에 따라 인권에 대한 간섭의 정당화 수준과 실사 요구사항의 엄격성이 달라질 수 있다는 것이다. 기본권 알고리즘영향평가는 심각도를 다음 세 가지 등급으로 구분하였으며, 심각도 등급에 따라 (7.4) 이하 부과되는 요구사항 또한 차등 적용될 수 있다고 보았다. 기본권 침해가 기본권의 핵심에서 멀어질수록 침해가 덜 심각하고, 보다 온건한 실사 및 정당화 요구가 이루어진다.

[그림 15] 네덜란드 기본권 알고리즘영향평가 심각도 등급



나. 평가의 기준

기본권 알고리즘영향평가는 알고리즘 시스템의 개발 또는 조달을 검토 중인 공공기관이 알고리즘 계획의 초기 단계에서 인권과 관련된 쟁점을 검토할 수 있도록 구조화된 질의를 제공한다. 이 질의는 공공기관 의사결정자가 대화 지향적이고 질적인 접근 방식을 사용하여 이해관계자와 협의하고 다양한 절충안을 고려하여 알고리즘 시스템의 개발 및 조달의 진행 여부와 방법에 대하여 적절한 결정을 내릴 수 있도록 지원한다.[92)]

다. 다른 규제 메커니즘과의 관계

기본권 알고리즘영향평가는 각 질의에서 관련되어 있거나 참조할 수 있는 다른 지침 또는 개인정보보호 영향평가 등 평가도구를 제시하면서 포괄적인 평가도구를 지향하였고, 기본권 알고리즘영향평가의 답변과 타 평가의 답변을 상호 활용하도록 하였다.

라. 운영상 쟁점

1) 평가 수행 주체

기본권 알고리즘영향평가는 각 공공기관에서 수행할 것을 예정하고 있다. 의회에서는 공공기관 의무화를 요구하였으나 2022년 현재로서는 자율적인 규범이다. 기본권 알고리즘영향평가는 평가 기준이 될 수 있는 질의 도구 기능에 우선 충실하고자 하였다. 따라서 평가 수행 주체의 독립성 등 운영상 쟁점에 대하여는 자세하게 언급하고 있지 않다.

2) 이해관계자 참여

기본권 알고리즘영향평가는 유의미한 평가를 위하여 단계별로 다양한 이해관계자가 참여하여 협의할 것을 권장하였으며, 여기에는 내부 이해관계자와 외부 이해관계자가 모

92) Open Government Partnership(2022). Algorithms and Human Rights: Understanding Their Impacts. <https://www.opengovpartnership.org/stories/algorithms-and-human-rights-understanding-their-impacts/(접근일: 2022. 8.15)>.

두 포함된다. 기관 내부 또는 직접 관련된 이해관계자로는 기관의 관리직, 최고정보책임자(CISO/CIO), 소통 전문가, 데이터 과학자, 개인정보 처리자 및 보유자, 해당 분야 전문가(직원), 개인정보보호 담당관, 인사팀 직원, 법률 자문, 알고리즘 개발자, 평가 의뢰인, 사업팀원, 사업팀장, 전략적 윤리 자문 등이 있으며, 외부 이해관계자로는 이익단체, 시민 패널 등이 있다.

3) 평가 결과 공개

질의 도구 기능에 충실한 기본권 알고리즘영향평가는 알고리즘영향평가 결과물의 공개 등 운영상 쟁점에 대하여는 자세하게 언급하고 있지 않다. 다만 인권 침해의 심각도 설정 등 쟁점을 검토할 때 공개적인 방식으로 논의할 것을 권장하고 있다.

5. 주요 빅테크 기업의 인권영향평가

유엔 기업과 인권 이행지침은 모든 기업에 자신의 활동이 인권에 부정적 영향을 초래하거나 이에 기여하지 않도록 하며, 부정적 영향이 발생할 경우 이에 대처할 책임을 부여하고 있다. 그리고 인권영향평가는 유엔 기업과 인권 이행지침에서 권고하는 인권실사 의무의 핵심 도구이다. 따라서 인공지능 인권영향평가를 따로 규정하지 않더라도, 인공지능을 개발하거나 도입하는 기업은 해당 인공지능이 인권에 미치는 영향을 판단하기 위하여 인권영향평가를 활용할 수 있다. 아직 인공지능 인권영향평가가 제도화되지 않았음에도, 주요 빅테크 기업들은 자사의 인공지능 혹은 알고리즘에 대해 인권영향평가를 수행한 바 있다. 이들 기업들도 인공지능 인권영향평가를 정례화한 절차를 두지 않았을뿐더러 때로는 형식적이라는 비판도 받고 있지만, 이들이 공개한 보고서를 통해 인공지능 인권영향평가가 어떠한 의미가 있는지 참고할 수 있다.

가. 구글의 유명인 인식 API에 대한 인권영향평가

구글은 기업 고객이 구글의 유명인(celebrity) 이미지 데이터베이스를 사용하여 자신의

콘텐츠에서 프레임이나 씬 단위로 유명인을 식별할 수 있는 API를 개발하면서 인권영향평가를 수행하였다. 정의롭고 지속가능한 세계의 형성을 위해 사업체와 협력하는 것을 목표로 하는, 기업들의 연합체 조직인 BSR에 인권영향평가를 의뢰[93]하였고, 2019년 10월 인권영향평가 요약 보고서[94]가 공개되었다.

BSR은 유엔 기업과 인권 이행지침에 기반한 방법으로 인권영향평가를 수행했다고 한다. 즉, 인권영향평가 과정에서 영향을 받는 이해관계자와 협의를 하였고, 독립적인 전문가와 대화하였으며, 취약성이 큰 그룹에 특별한 관심을 기울였다. 평가 과정에서 구글 클라우드 AI의 API 제품팀 및 AI 원칙팀과 협력하였다. 구글이 재정지원을 하였지만, BSR은 보고서 내용에 대해 통제권을 보유하여 독립성을 보장받았다고 한다.

유명인 인식 API와 관련하여, BSR과 구글이 식별한 인권 위험은 다음과 같다. 첫째, "유명인"에 대한 합의된 정의가 없기 때문에, 데이터베이스에 포함되는 사람을 결정하는 데 있어, 예를 들어 아동권에 영향을 미칠 수 있다. 둘째, 어떻게 의미있는 동의 문제를 획득할 것인가의 문제가 있는데, 당사자 의사에 반해 데이터베이스에 포함될 경우 프라이버시권을 침해할 수 있기 때문이다. 셋째, 유명인도 취약 그룹(vulnerable group)일 수 있는데, 이들의 유명세가 나이, 젠더, 정치적 신념, 종교, 인종 등과 결합될 경우 일반인보다 대규모로 차별, 괴롭힘, 혐오 발언에 노출될 수 있기 때문이다. 또한 유명인에 대한 콘텐츠가 폭력이나 괴롭힘을 야기하는 방식으로 딱지붙이거나 가짜뉴스와 결부될 경우 보안 위험을 야기할 수도 있다. 넷째, 인권 위험은 콘텐츠에 따라 다른데, 통상 주류 미디어 영상보다 이용자 콘텐츠나 CCTV 영상 콘텐츠 등이 더 위험하다.

이러한 위험을 식별한 후에 BSR은 부정적 영향을 방지, 완화할 수 있는 정책을 권고했고, 구글은 이를 API 개발에 반영하였다. 이 중 일부는 다음과 같다. 첫째, 유명인 인식 API의 사용을 기업 소비자가 소유했거나 사용허락을 받은 전문 영상 미디어 콘텐츠(즉, 이용자 제작 콘텐츠가 아닌 콘텐츠)로 제한하는 '서비스 특별 약관'을 적용할 것, 둘째, 자발적으로 공중 미디어 관심사의 대상이 된 개인으로 유명인 데이터베이스를 제한할 것, 셋째, 유명인이 요청할 경우 구글이 관리하는 유명인 데이터베이스에서 제외하

93) BSR(2019). Google's Human Rights by Design. <https://www.bsr.org/en/blog/google-human-rights-impact-assessment-celebrity-recognition(접근일: 2022. 11. 1)>.

94) BSR(2019).

는 옵트아웃 정책을 구현할 것, 넷째, 자격있는 산업(엔터테인먼트, 미디어, 스포츠) 영역 내에서, 허용되는 사용 사례를 선언하고, 전문적인 영상 미디어로 사용을 제한하는데 동의한 기업 고객으로 고객 '화이트리스트'를 구축할 것 등이다.

이와 더불어 BSR은 구글 뿐만 아니라 유명인 인식 도구를 제공하는 다른 제공자와 미디어 및 엔터테인먼트 산업에 대해서도 다음과 같이 권고하고 있다. 구글은 유명인 인식 도구를 제공하는 하나의 업체에 불과하기 때문에, BSR은 구글 뿐만 아니라 산업계 차원의 정책이나 표준이 만들어질 필요가 있다고 제안한다. 또한 미디어 및 엔터테인먼트 업체의 역할도 강조하면서, 유명인 인식 도구 및 다른 얼굴인식 도구를 사용하는 업체들이 인권실사를 이행할 것을 권고한다.

요약 보고서는 인권영향평가 방법론을 자세하게 언급하고 있는 것은 아니지만, 인권영향평가가 기업의 제품 설계나 정책에 어떠한 영향을 미칠 수 있는지 잘 보여준다.

나. 페이스북의 국가별 인권영향평가

2018년 11월 5일, 페이스북은 미얀마에서 페이스북의 역할에 대한 인권영향평가 보고서[95]를 발표하였다. 이 인권영향평가 역시 BSR에 의해 수행되었다.[96]

이 보고서는 2018년 이전 페이스북이 자신의 플랫폼이 분열과 폭력의 조장을 방지하는데 충분한 역할을 하지 못했다고 결론을 내렸다. 2018년에는 페이스북이 미얀마에서의 페이스북의 남용을 막기 위해 인력, 기술, 파트너쉽에 더 많은 투자를 했으며 보고서는 이러한 시정 조치들을 인정한다고 밝혔다.

보고서에서 BSR은 5가지 권고안을 제시했는데, 이는 첫째, 거버넌스 및 책임성 구조에 기반할 것, 둘째, 콘텐츠 정책의 집행을 증진할 것, 셋째, 지역 이해관계자의 참여를 강화할 것, 넷째, 규제 개혁을 옹호할 것, 다섯째, 미래를 준비할 것 등이다.

거버넌스와 관련하여 BSR은 페이스북에 독자적인 인권 정책의 채택, 인권 전략을 감독할 공식적 거버넌스 구조의 수립, 정기적인 업데이트를 권고했다. 페이스북은 자신들

95) BSR(2018).

96) Meta(2018. 11. 5). An Independent Assessment of the Human Rights Impact of Facebook in Myanmar. <https://about.fb.com/news/2018/11/myanmar-hria/(접근일: 2022. 11. 1)>.

의 플랫폼 정책이 국제인권기준에 기반하여 수립되었다고 주장하며, BSR의 권고에 따라 인권 측면에서 콘텐츠 관리 정책을 들여다보고 있다고 말했다.

콘텐츠 정책의 집행과 관련해서는, BSR은 플랫폼에서 무엇이 허용되고 무엇이 허용되지 않는지에 대한 정책의 집행을 증진할 것, 특히 미얀마의 지역적 맥락을 이해하는 팀을 개발할 것을 권고했다. 이에 페이스북은 미얀마 전담팀을 구성하고 증오발언 탐지 문제를 개선했다고 밝혔다.

그러나 하버드 케네디스쿨의 인권정책을 위한 카센터(CARR Center)의 한 연구자는 페이스북의 미얀마에서의 인권영향평가가 인권영향을 제대로 평가하지 못했다고 비판한다.[97] 그 이유는 첫째, 이 인권영향평가가 페이스북 뉴스피드 알고리즘의 인권영향에 대해 다루지 않았다는 것이다. 페이스북 알고리즘은 이용자의 관심을 가장 잘 포착할 수 있다고 판단되는 포스팅에 우선순위를 두는데 이는 자극적인 가짜뉴스와 허위정보의 확대에 기여했을 수 있는데 이에 대해 평가하지 않은 것이다. 둘째는 인권영향평가가 미얀마 역사의 일부를 다루기는 했지만, 이러한 맥락의 핵심적 요소들이 미얀마 뉴스 피드와 같은 페이스북 제품의 보급 및 운영 결정과 관련되는지에 대해 다루지 않았다는 것이다. 미얀마 정부의 로힝야 탄압의 역사 역시 거의 다루지 않았다. 이와 같은 분석의 문제점 때문에 인권영향평가는 페이스북이 남용된 이유를 미얀마의 사회, 문화, 정치적인 환경 탓으로 돌릴 뿐, 미얀마에서의 인권 침해를 야기하거나 기여한 것에 대한 페이스북의 책임을 제대로 다루지 않았다는 것이다.

2020년 5월 12일, 페이스북은 2018년에 시행한, 스리랑카, 인도네시아, 캄보디아에서의 페이스북 서비스의 역할을 평가한 인권영향평가 보고서와 이에 대한 페이스북의 대응 계획을 발표하였다. 캄보디아의 평가는 BSR이, 스리랑카와 인도네시아는 Article One이 담당하였다.[98]

2019년 말, 페이스북은 인도의 페이스북 서비스에 대한 인권영향평가를 Foley Hoag라

97) CARR Center(2021. 3. 19). Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar. <https://carrcenter.hks.harvard.edu/publications/human-rights-impact-assessments-ai-learning-facebook%E2%80%99s-failure-myanmar(접근일: 2022. 11. 1)>.

98) Meta(2020. 5. 12). An Update on Facebook's Human Rights Work in Asia and Around the World. <https://about.fb.com/news/2020/05/human-rights-work-in-asia/(접근일: 2022. 11. 1)>.

는 로펌에 의뢰하였다. 2019년에 시민사회단체들이 인도에서 페이스북의 콘텐츠 정책과 콘텐츠 관리 절차를 비판하는 몇 개의 보고서를 발표했기 때문이다. 그런데 인도 인권영향평가 보고서는 2022년 상반기까지 발표되지 않았으며, 2021년 11월 월스트리트저널은 페이스북의 인권팀이 독립적인 인권영향평가의 범위를 축소하려는 움직임을 보였다고 보도했다.[99] 인권단체들은 이 보도를 둘러싸고 여러 우려를 나타냈는데, 인도에서 페이스북을 통해 무슬림과 다른 소수자 그룹을 대상으로 한 혐오발언과 가짜뉴스가 난무했음에도 불구하고, 페이스북은 이러한 콘텐츠를 삭제하기 위한 조치를 충분하게 취하지 않았기 때문이다.[100] 또한 월스트리트저널은 인도 페이스북과 인도 집권당 사이의 결탁 관계에 대해 보도하기도 했다.[101]

2022년 7월, 메타는 2020-2021년 활동을 다룬 메타 인권보고서를 발행했는데, 여기에 인도 인권영향평가 보고서의 내용을 4p로 간단하게 포함하였다.[102] 그러나 메타는 시민사회로부터 인도 인권영향평가 보고서를 은폐하려 한다고 비판을 받았다. 전체 보고서뿐만 아니라 보고서의 권고가 무엇인지도 자세한 내용이 공개되지 않았기 때문이다.[103]

메타의 인권영향평가 사례는, 인권영향평가가 자칫하면 기업으로 하여금 인권 보호 책임을 충족하는 외양만 갖추게 할 뿐, 실질적으로는 그 책임을 회피하는 수단으로 전락할 수 있음을 보여준다. 또한, 인권영향평가의 독립적 수행과 보고서 공개를 통한 투명성

---

99) The Wall Street Journal(2021. 11. 12). Facebook Is Stifling Independent Report on Its Impact in India, Human Rights Groups Say. <https://www.wsj.com/articles/facebook-is-stifling-independent-report-on-its-impact-in-india-human-rights-groups-say-11636725601(접근일: 2022. 11. 1)>.

100) APC(2022. 1. 21). Release of the Human Rights Impact Assessment of Facebook in India. <https://www.apc.org/en/pubs/release-human-rights-impact-assessment-facebook-india(접근일: 2022. 11. 1)>.

101) The Wall Street Journal(2020. 8. 14). Facebook's Hate-Speech Rules Collide With Indian Politics. <https://www.wsj.com/articles/facebook-hate-speech-india-politics-muslim-hindu-modi-zuckerberg-11597423346(접근일: 2022. 11. 1)>.

102) Meta(2022). Meta Human Rights Report : Insights and Actions 2020-2021. July 2022. <https://about.fb.com/wp-content/uploads/2022/07/Meta_Human-Rights-Report-July-2022.pdf(접근일: 2022. 11. 1)>.

103) Time(2022. 7. 14). Facebook Accused of 'Whitewashing' Long-Awaited Human Rights Report on India. <https://time.com/6197154/facebook-india-human-rights/(접근일: 2022. 11. 1.)>; Access Now(2022. 7. 29). Meta must disclose India's Human Rights Impact Assessment. <https://www.accessnow.org/meta-india-human-rights-impact-assessment/(접근일: 2022. 11. 1)>.

확보가 인권영향평가의 신뢰성을 위해 매우 중요하다는 점을 알 수 있다.

다. 마이크로소프트의 책임있는 인공지능 인권영향평가 가이드

2022년 6월, 마이크로소프트(MS)는 <책임있는 인공지능 영향평가 가이드>를 발간하였다. 이 가이드는 평가 템플릿을 포함하고 있으며, MS는 이 가이드를 공개하여 자신들의 성과를 공유하고 다른 사람들의 의견을 받으며, 인공지능을 둘러싼 더 나은 규범과 관행에 기여하고자 하는 목적이라고 밝히고 있다.[104)]

이 가이드는 서로 다른 전문성을 가진 사람으로 평가팀을 구성하고 템플릿에 구성원의 토론 내용을 기록하도록 하고 있다. 템플릿의 내용은 이후 잠재적인 평가자가 검토하는 관점에서 작성된다. 또한 이 가이드는 영향평가의 각 단계에 맞춰 병원에서 자원 관리를 위한 인공지능 시스템을 도입하는 것을 사례로 제시하고 있다.

이 가이드는 다음과 같이 구성되어 있다.

○ 섹션 1. 프로젝트 개요 : 시스템 프로파일, 시스템의 생명주기 단계, 시스템에 대한 설명, 시스템의 목적, 시스템의 특성, 지리적 영역 및 언어, 배포 모드(온라인 서비스로 배포되는지, 코드 형태인지 등), 시스템 사용의 식별(가능한 사용방식에 대한 브레인스토밍, 가능한 사용을 의도된 것, 지원되지 않는 것, 오용으로 분류, 어떠한 사용이 민감한 상요 혹은 제한된 사용인지 점검)

○ 섹션 2. 의도된 사용 : 목적 적합성 평가, 이해관계자의 잠재적 이익 및 해악, 책임있는 인공지능 표준의 목표 중심 요구사항을 위한 이해관계자, 공정성 고려사항, 기술 준비도 평가, 업무 복잡성, 인간의 역할, 배포 환경 복잡성

○ 섹션 3. 부정적 영향 : 제한된 사용, 지원되지 않는 사용, 알려진 제한사항, 실패의 이해관계자에 대한 잠재적 영향, 오용의 이해관계자에 대한 잠재적 영향, 민감한 사용

○ 섹션 4. 데이터 요구사항 : 데이터 요구사항, 기존 데이터셋

○ 섹션 5. 영향 요약 : 잠재적인 해악과 예비적 완화조치, 목표 적용가능성, 영향평가 서명

104) Microsoft(2022).

제3절 국내 인공지능 기준과 인권영향평가

인공지능이 사회에 미치는 영향에 대한 우려가 커짐에 따라 우리나라에서도 개인정보보호위원회, 과학기술정보통신부, 금융위원회, 서울특별시교육청 등 중앙부처 및 지방자치단체가 인공지능의 분야별 위험을 점검하기 위한 기준 및 도구를 보급하여 왔다. 현재까지 이들 도구들은 자율점검 절차로 사용되고 있다.

한편, 유엔 기구들의 인권실사 권고가 계속되고 최근 유럽과 미국 등 주요국가 중심으로 인권경영이 제도화되어 옴에 따라, 국내에서도 공공기관·공기업과 상장기업 우선으로 인권영향평가의 실시와 보고가 이루어지기 시작하였다. 국가인권위원회, 법무부, 산업통상자원부는 기관과 기업의 인권영향을 평가할 수 있는 도구와 정책을 발표하여 왔다.

이하에서는 먼저 국내에 보급된 인공지능 기준들을 살펴본 후, 국가인권위원회의 인권영향평가 정책을 중심으로 인공지능 인권영향평가의 적용가능성을 살펴본다.

1. 인공지능 자율점검 기준

개인정보보호위원회는 인공지능 챗봇 이루다의 개인정보 보호법 위반 사건[105]이 발생한 후, 2021년 5월 <인공지능(AI) 개인정보보호 자율점검표>를 발표하였다.[106] 이 점검표는 인공지능의 수명주기를 ①인공지능의 기획·설계, ②개인정보 수집, ③개인정보 이용·제공, ④개인정보 보관·파기, ⑤인공지능 서비스 관리·감독, ⑥인공지능 서비스 이용자 보호 및 피해구제, ⑦개인정보 자율보호 활동, ⑧인공지능 윤리 점검 단계로 구분하고 인공지능의 개발·운영에 참여하는 자가 각 단계에서 준수해야 할 개인정보 보호법의 기준을 체크리스트 형식으로 제시하였다.

과학기술정보통신부는 2021년 5월 <신뢰할 수 있는 인공지능 실현전략>[107]을 발표하면서, 소관 「지능정보화 기본법」 제56조에 기반하여 “인공지능이 국민생활 전반에 미치

105) 개인정보보호위원회(2021. 4. 29). 개인정보위, '이루다' 개발사 (주)스캐터랩에 과징금·과태료 등 제재 처분; 개인정보보호위원회 2021. 4. 28. 결정 제2021-007-072호 심의·의결서.

106) 개인정보보호위원회(2021. 5. 31). 개인정보위, 인공지능(AI) 자율점검표 발표; 개인정보보호위원회(2021).

107) 과학기술정보통신부(2021. 5. 14).「신뢰할 수 있는 인공지능 실현전략」발표.

는 영향을 체계·종합적으로 분석하고 대응하기 위해” 인공지능 영향평가를 실시할 방침임을 밝혔다. 한편 2022년 과학기술정보통신부는 한국정보통신기술협회(TTA)와 함께 개발한 <2022 신뢰할 수 있는 인공지능 개발 안내서(안)>을 공개하였다. 이 안내서(안)은 인공지능 서비스 및 제품 개발 업무에 종사하는 개발자가 인공지능의 신뢰성을 제고하기 위해 참고할 수 있는 자료로서 편찬되었으며, 업무 환경과 상황, 그리고 개발 목적을 고려하여 안내서의 내용을 필요한 대로 선택하여 활용할 것을 권장하였다. 과학기술정보통신부는 2022년 3월 이 안내서(안)의 요구사항과 검증항목을 인증하는 체계를 구축하고 시범대상 기업에 적용하는 사업을 추진하기 시작했다.[108] 이 인증 체계는 민간 자율적으로 자사 제품의 신뢰성 관련 중요 정보를 온라인에 공시하는 기준으로 추진되고 있다.

금융위원회는 2021년 7월, <금융분야 인공지능(AI) 가이드라인>의 시행을 발표하였다.[109] 이 가이드라인은 금융회사가 인공지능 기반 금융서비스를 개발 및 활용할 때 신뢰 제고에 필요한 준칙을 제시하였다. 금융위원회는 가이드라인에 대한 금융회사 실무자들의 의견을 수렴한 결과를 토대로 2022년 8월 <금융분야 AI 개발・활용 안내서>를 발표하였다.[110] 이 안내서는 앞서 가이드라인의 규제 불확실성을 해소하는 한편, 인공지능 활용 과정에서 발생 가능한 위험을 예방・관리하는 데 목적이 있다. 안내서는 가이드라인에서 제시한 1) 목적과 적용 범위, 2) 거버넌스의 구축, 3) 기획・설계 단계, 4) 개발 단계, 5) 평가・검증 단계, 6) 도입・운영・모니터링 단계, 7) AI 업무위탁에 대한 특례 항목을 재구성하여, 금융회사에 공통으로 적용되는 항목과 신용평가 및 여신심사, 이상거래 탐지, 챗봇, 맞춤형 상품 추천, 로보어드바이저 서비스의 5대 서비스별로 적용되는 항목을 구분하여 체크리스트를 제시하였다.

2021년 9월 서울특별시교육청은 <인공지능(AI) 공공성 확보를 위한 현장 가이드라인>을 발표하였다.[111] 이 가이드라인은 학교에서 인공지능을 도입할 때 ‘인공지능 등급 평가 매트릭스’를 사용해 인공지능의 영향을 평가하도록 하였다. 이 매트릭스는 ‘의사결

108) 과학기술정보통신부(2022. 3. 4). 2022년 과학기술정보통신부「민간 인공지능 신뢰성 시범인증」사업자 모집 공고. 과학기술정보통신부공고 제2022-0275호; 과학기술정보통신부(2022).

109) 금융위원회(2021. 7. 8).「금융분야 인공지능(AI) 가이드라인」이 시행됩니다: 금융권 AI 활용을 활성화하고 AI 기반 금융서비스에 대한 신뢰를 제고하기 위한 모범규준 마련·발표.

110) 금융위원회(2022. 8. 4). 금융분야 인공지능 활용 활성화 간담회 개최: 금융분야 인공지능 활용 활성화 및 신뢰확보 방안 발표; 금융위원회(2022).

111) 서울특별시교육청(2021. 7. 30). 서울시특별시교육청, 인공지능(AI) 공공성 확보를 위한 현장 가이드라인 공청회 개최; 서울특별시교육청(2021).

정 영향 정도’를 한 축으로 하여 낮음, 중간, 높음을 평가하고, 다른 축은 ‘개인정보 민감 정도’를 낮음, 중간, 높음으로 평가하여 2차원으로 나타나는 인공지능의 영향을 1등급에서 4등급까지 구분하였다. 이때 평가용 도구로는 △일반 △데이터 △알고리즘 및 △개인정보 항목에 대하여 총 10개 질의를 갖춘 ‘인공지능 영향평가 체크리스트’를 사용하도록 하였다. 평가 결과 영향 정도가 낮은 인공지능(4등급)은 단위학교에서 바로 사용이 가능하고, 영향이 높을수록 교내 AI위원회(3등급) 또는 외부위원을 포함하는 학교 AI위원회(2등급)에서 심의 및 조치를 거쳐 사용하도록 하였고, 영향이 가장 높은 인공지능(1등급)은 교육청 AI위원회 심의를 거쳐 채택 또는 조치하도록 하였다.

국내에서 인공지능에 대한 구속력 있는 영향평가 제도가 도입된다면 상기 기준들이 그 평가도구에 반영될 수 있을 것이다. 다만, 이 기준 및 도구들은 인공지능이 미치는 영향 중에 각 부처 소관사항을 살펴볼 뿐, 인권에 미치는 부정적인 영향을 식별하고 방지 및 완화하는 것을 주요 목적이나 내용으로 삼고 있지 않다.

2. 인권경영과 인공지능 인권영향평가

가. 인권실사 현황과 기준

인권경영이란 기업이 인권침해를 일으키거나 연루될 위험을 사전에 예방하고, 인권침해가 일어난 때에는 사후적으로 피해자를 구제할 수 있도록 기업의 인권존중 문화를 정착시키고 인권 중심 의사결정체계를 구축하는 경영활동을 말한다. 이때 인권실사는 기업이 끼치는 부정적 인권영향을 식별하고 방지·완화하는 절차로, 인권영향평가 실시, 인권영향평가 결과를 기업활동 전반에 반영·실천하는 조치, 조치의 효과성 모니터링 등을 포함한다.[112] 여기서 인권영향평가는 인권실사의 핵심적인 요소이다.

2011년 유엔 기업과 인권 이행지침 이후 인권경영이 국제규범으로 정착되었고, 프랑스 및 독일 등 주요 국가들은 ‘실사의무화법’을 제정하여 인권경영 실사제도의 법제화를 이루어 왔다. 유럽연합 집행위원회는 2022년 2월 23일 「기업 지속가능성 실사법(Directive)」을 채택하고 회원국 전체에 인권경영 실사제도의 이행을 요구하고 있다.

112) 국가인권위원회(2020. 5. 26). 인권위-법무부, 인권경영 확산 위해 손 맞잡는다.

유엔 인권기구는 한국 정부에 대하여, 기업의 인권경영 실천을 위한 지침을 제공하고 관련 법·정책을 도입할 것을 꾸준히 권고해 왔다. 그러나 우리나라에서 인권실사 및 인권영향평가는 아직까지 법률로 규정되어 있지 않다. 다만 정부가 2021년 12월 30일 국회에 발의한 인권정책기본법안에서 '제5장 기업과 인권'은 유엔 기업과 인권 이행지침의 구조와 개념을 수용하였다.[113] 당초 이 법 입법예고안(법무부공고제2021-198호)은 기업의 인권존중을 증진하기 위하여 "정부는 기업의 인권존중책임에 대한 평가기준 및 평가지표를 설정하여 운영할 수 있다(입법예고안 제23조 제3항)"고 규정하였으나, 이후 발의안에서는 정부가 "기업의 인권존중책임 실천을 위한 세부 지침"과 "기업의 인권존중책임 실천 관련 정보의 자율적 공개를 위한 표준"을 마련하여 보급하고 그 활용을 장려한다고 규정하였다는 차이가 있다(발의안 제18조제2항).

이와 같은 상황에서 현재 우리나라 인권영향평가는 인권실사를 의무화해 온 국제 기준의 압력을 받은 경영평가제도를 통하여 간접적으로 의무화되고 있다. 특히 상장기업[114] 우선으로 환경·사회·지배구조(Environment, Social, Governance, 일명 'ESG') 공시가 추진되면서 기업과 그 사업에 대한 인권영향평가가 도입되고 있다. 한편, 기획재정부는 2018년도 <공공기관 경영평가편람>에서 윤리경영 지표에 인권 항목을 포함한 데 이어, 2022년 2월 4일 「공공기관의 통합공시에 관한 기준」을 개정하여 기관운영 대항목 하에 인권경영 항목을 독립적으로 신설 및 공시하도록 하였다.[115] 행정안전부는 2019년 5월에 발표한 <지방공기업 경영평가편람>에서 인권경영을 윤리경영 항목에서 독립된 별도지표로 신설한 바 있다. 이러한 배경에서 공공기관과 공기업에도 인권영향평가의 실시 및 공개가 확산되어 왔다.

국가인권위원회와 각 정부부처는 인권경영에서 인권영향을 평가하기 위한 도구들을 개발하고 보급하여 왔다. 법무부는 2019년 5월 <인권경영 표준지침(안)>을 발표하였고, 산업통상자원부는 2021년 12월 <K-ESG 가이드라인 v1.0>을 발표하였다. 특히 국가인권위원회는 2014년 <인권경영 가이드라인 및 체크리스트>와 2018년 <공공기관 인권경영

113) 김동현(2022), 110면.

114) 국내에서는 2025년부터 단계적으로 ESG 공시가 의무화되고, 2030년에 전체 코스피 상장사 대상으로 확대될 예정이다. 한국경제신문(2021. 11. 30). 내년부터 상장 심사 때 기업 'ESG 체력' 검증; ESG경제(2022. 7. 8). 한국거래소, "ESG 공시 '사업보고서'에 통합할 필요 없어".

115) 기획재정부(2022. 2. 7). 통합공시를 통해 공공기관의 ESG 경영 선고: 공공기관 알리오(Alio) 통합공시 기준 개정.

매뉴얼>을 발표하고 공공기관장들에게 그 적용 및 활용 등을 권고하여 왔고, 현재 대부분의 공공기관은 이 기준과 도구를 참고하여 인권경영체계를 구축하고 있다.[116]

국가인권위원회 <공공기관 인권경영 매뉴얼>이 권고하고 있는 인권영향평가 추진체계 및 도구에 대하여 살펴보면 다음과 같다. 우선 인권경영 대상 기관은 인권경영 담당 부서와 담당자를 지정하고 인권경영위원회를 구성하여야 한다. 인권경영 담당부서는 인권영향평가 실시 계획 및 체크리스트를 마련하는 한편 기관 내부 교육을 실시하는 등 인권영향평가 절차를 주무하는 역할을 한다. 인권영향평가를 실시하는 인권경영위원회는 임직원, 노동조합, 공급망, 지역주민, 고객, 인권전문가 등 기관 내외부 이해관계자로 구성된다.

인권경영 평가도구로는 국가인권위원회 가이드라인 및 체크리스트 지표 등을 기반으로 각 기관의 실정에 맞는 체크리스트를 마련할 것이 권장된다. 특히 사업 인권영향평가의 경우 인권경영 담당부서가 먼저 사업부서와 이해관계자로부터 다양한 정보와 의견을 수렴하여 대상 사업의 실제적·잠재적 인권위험을 분석하고, 이렇게 분석한 위험을 바탕으로 지표를 선정하고 인권영향평가 체크리스트를 구성한다.

체크리스트는 일반적으로 '예', '보완필요', '아니오', '정보 없음', '해당 없음' 등의 응답이 가능하도록 구성할 수 있으나, 기관에 따라 다양한 형태로 답변 결과를 제시할 수 있다. 체크리스트에 사용할 문장은 인권 위험을 함축적으로 나타내는 문장으로 표현하되, 평가자에 따라 중의적으로 해석되는 일이 없도록 쉽고 간결한 문장을 사용한다. 지표가 국내·외 법률 등 규범과 관련되어 있을 경우에는 해당 규범을 그대로 활용하여 작성할 수 있으며, 특정 인권 위험에 대해 단답 형태로 평가하기 어려운 경우 서술형 평가를 일부 병행할 수 있다. 체크리스트 초안 작성을 완료한 이후 기관 내·외부 의견수렴을 거쳐 체크리스트를 확정한다.

각 사업부서는 가이드라인과 체크리스트에 대한 교육을 받고 인권영향평가 세부평가지표에 맞는 증빙자료를 제출하며, 인권경영 담당부서는 취합된 자료를 인권경영위원회에 제출한다. 인권경영위원회는 체크리스트를 활용하여 취합된 자료에 대한 평가를 실시하고, 인권영향평가 결과를 인권경영 담당부서에 제출하는데, 이때 평가결과는 체크리스

116) 국가인권위원회 결정 2022. 5. 4. 공공기관·공기업 인권경영 강화를 위한 인권경영 보고 및 평가 지침 적용 권고.

트 평가 결과, 이해관계자 의견, 실제적·잠재적 인권 위험에 대한 방지 조치 등을 포함하여야 한다. 인권경영 담당부서는 평가결과를 최고경영진에 보고하고, 최고경영진은 인권침해 방지 조치를 수립·시행하며, 지속적인 모니터링을 실시한다. 더불어 인권영향평가 결과는 이해관계자들에게 홈페이지 등을 통해 공개된다.

한편 국가인권위원회는 평가의 객관성과 신뢰성을 높이기 위해, 2022년 7월 13일 <인권경영 보고 및 평가 지침>을 발표하고, 정부 부처 및 광역자치단체장에 대하여 산하 공공기관이 이 지침에 따라 독립적인 항목으로 인권경영을 평가할 것을 권고하였다.[117] 국가인권위원회는 각 기관마다 인권경영의 평가방법, 세부평가항목 및 배점이 다르고 평가기준도 모호하여 인권경영 평가 결과의 신뢰성 및 타당성에 의문이 제기될 소지가 큰 상황이라고 밝혔다. 또한 공공기관의 90% 이상이 사업 후에 인권영향평가를 수행하고 있으며, 주요사업 인권영향평가에 대한 이행률이 50%를 밑도는 등 상당수 기관이 인권경영 실현의 한계에 봉착해 있는 것으로 조사되었다. 이에 지침은 인권경영의 '보고지침'과 '평가지침'을 각기 제시하였다. 인권경영을 실천해야 하는 공기업·공공기관을 우선 대상으로 하였지만, 업종이나 규모에 관계 없이 모든 기업이 이용할 수 있도록 하였다.

인권영향평가와 관련한 보고지침으로는 1) 해당연도 인권영향평가 실시과정, 2) 식별된 주요 인권이슈, 3) 주요 인권이슈에 대한 대응계획, 4) 주요 인권이슈의 대응계획 성과를 반드시 보고하도록 하였다. 특히 지침은 인권경영 보고에서 '주요 인권이슈'를 반드시 명시할 것을 요구하며, '중대성 평가'를 통해 기관 및 기업이 해당년도에 대응하기로 결정한 주요 인권이슈가 인권영향평가의 최종결과물로서 도출되어야 한다고 강조하였다. '중대성 평가'는 식별된 여러 인권이슈 중에서 관련 실태조사, 이해관계자의 복잡성과 다양성, 인권영향 수준, 위법여부, 침해의 심각성(예컨대 생명·신체와 관련한 침해는 재산에 대한 침해보다 더 심각하다고 할 수 있다), 피해자의 수, 피해의 사후적 구제가능성, 국가 정책, 기업의 업종, 피해 발생 장소 등 다양한 변수를 고려하여, 우선적으로 대처할 '주요 인권이슈'를 도출하는 것이다. 주요 인권이슈를 도출하지 못한 인권영향평가는 제대로 된 인권영향평가로 간주되기 힘들다.

---

117) 국가인권위원회 보도자료(2022. 8. 16). 공공기관의 인권경영 강화를 위한 '인권경영 보고 및 평가지침'적용 권고.

인권경영 평가지침의 경우, 전체 경영평가 중 인권경영 평가 배점을 5점으로 가정하였을 때, △인권경영체계 및 인권경영 정책(최대 1점) △인권영향평가(최대 2점) △구제절차(최대 1점) △인권경영의 소통 및 인권경영 교육(최대 1점)을 항목별로 평가하도록 하였다. 인권영향평가에 대한 경영평가는 주요 인권이슈를 도출했는지, 그에 대한 대응계획과 실행이 적정했는지를 평가한다.

인공지능 인권영향평가 기준과 도구를 수립함에 있어, 이와 같은 인권영향평가에 대한 국가인권위원회 기준을 이후 살펴볼 인공지능에 대한 국가인권위원회 가이드라인과 더불어 고려할 필요가 있다.

나. 인공지능 인권 가이드라인과 인권영향평가

국가인권위원회는 2022년 5월 17일 <인공지능 개발과 활용에 관한 인권 가이드라인(이하 '인공지능 인권 가이드라인')>을 공개하였다. 이 가이드라인은 인공지능 개발과 활용에 있어 △인간의 존엄성 및 개인의 자율성과 다양성 보장, △투명성과 설명 의무, △자기결정권의 보장, △차별금지, △인공지능 인권영향평가 시행, △위험도 등급 및 관련 법·제도 마련 등을 주요 내용으로 하였다. 국가인권위원회는 국무총리 및 관련 부처 장관·기관장에게 가이드라인에 기초하여 인공지능 관련 정책을 수립·이행하고 관계 법령을 제·개정할 것 등을 권고하였으며, 2022년 10월 21일 국무총리 등이 이 권고를 수용하였다고 발표하였다.[118]

가이드라인은 특히 인공지능의 개발과 활용에 있어 인권영향평가의 시행을 요구하면서, 인권침해와 차별의 가능성 및 정도, 영향을 받는 당사자의 수, 사용된 데이터의 양 등을 고려하여 공공기관 및 민간기업을 대상으로 인권영향평가를 실시하여야 한다고 하였다. 특히 기존 제도로 관리되거나 감독될 수 없는 새로운 분야일수록 인권영향평가 제도를 도입해야 한다. 인권영향평가 내용은 인공지능의 특성, 상황, 범위 및 목적을 감안하여 인권 가이드라인이 제시한 원칙 및 내용, 국제 인권 기준, 관련 법률에서 정한 의무 등을 포함하여야 하며, 인권침해 위험요인의 분석, 개선 사항 등을 도출해야 한다. 인

118) 국가인권위원회(2022); 국가인권위원회(2022. 5. 17). 인권위, <인공지능 개발과 활용에 관한 인권 가이드라인> 마련; 국가인권위원회(2022. 10. 21). 〈인공지능 개발과 활용에 관한 인권 가이드라인〉 권고, 국무총리 및 관련 부처 장관·기관장 수용.

권영향평가는 개발 및 출시 전에 실시하고 인공지능의 기능 또는 범위 변경 시 평가를 갱신하여야 한다.

국가와 기업은 인권영향평가 결과에서 인권에 미치는 부정적인 영향이나 편향성 및 위험성이 드러난 경우 이를 방지하거나 완화하기 위한 조치사항을 수립하여 적용하여야 하며, 원칙적으로 그 내용을 공개해야 한다. 또한, 이를 방지하거나 완화하는 조치를 취하기 전에는 그 개발과 활용을 중단해야 한다. 국가는 인권영향평가를 인권전문성과 독립성을 확보한 기관이 담당하도록 하고, 인권영향평가의 활성화를 위하여 관계 전문가의 육성, 영향평가 기준의 개발 및 보급 등 필요한 조치를 마련해야 한다.

국내에 도입되는 인공지능 인권영향평가는 국가인권위원회 가이드라인에서 제시하는 원칙에 부합하는 방향으로 도입되는 것이 바람직할 것이다. 더불어 국가인권위원회는 앞서 살펴본 대로 인권경영 확산을 위하여 인권영향평가의 기준과 지표를 꾸준히 제시하여 왔고, 인권영향평가를 실시하는 기관과 기업들은 보급된 평가도구들을 그대로 사용하거나 최적화하여 사용해 왔다.

인공지능 관련 사업을 운영하고 있거나 계획을 하고 있는 기관이나 기업의 경우 해당 사업에 대한 인권영향평가 또한 실시하여야 한다. 이들 기관이나 기업이 인공지능 사업에 대한 인권영향평가를 실시할 때 본 연구에서 개발한 도구를 비롯하여 국내외에서 제안되어 온 다양한 인공지능 평가 기준 및 도구들을 여타의 인권영향평가도구들과 결합하고 최적화하여 사용할 수 있을 것이다.

제4절 시사점

앞서 살펴보았듯이 인권 위험을 비롯하여 인공지능의 위험을 예방하고 완화하기 위하여 국내외에서 인공지능 영향평가를 다양하게 검토하고 제안하여 왔으며, 일부 국가에서는 이미 제도적 수준에서 이를 시행하고 있는 사례도 있었다. 다른 한편에서는 최근 인권실사와 그 핵심인 인권영향평가를 제도화하여 시행하기 시작한 국내외 동향을 볼 수 있었다. 이러한 흐름에 비추어 우리나라 인공지능 인권영향평가의 개발과 시행에 대하여 생각해볼 수 있는 시사점은 다음과 같다.

캐나다 알고리즘영향평가, 영국 조달지침 및 NMIP 알고리즘영향평가는 특히 공공부문에서 사용하거나 공공데이터를 활용하는 인공지능에 영향평가를 제도화하여 시행하고 있다. 민간까지 적용되는 인공지능 영향평가의 경우, 특히 인권에 위험한 영향을 미칠 우려가 큰 인공지능을 대상으로 제도화가 추진되고 있다. 유럽연합은 2021년 인공지능법(안)에서 고용이나 사회복지 등 고위험(high-risk) 인공지능시스템에 대하여 출시 전 위험의 평가, 제거 및 완화 조치를 시행하도록 하는 위험관리를 의무화하였다. 2022년 11월 시행된 디지털서비스법은 대규모온라인플랫폼에 대하여 연 1회 이상 또는 중대한 위험을 미치는 기능을 배치하기 전에 알고리즘 위험평가를 실시하도록 하였다. 미국 의회에 발의된 2022년 알고리즘 책무성법(안)은 일정 규모 이상 기업의 자동화된 의사결정 시스템 및 고용, 주택, 금융 등 중요 의사결정 프로세스를 대상으로 알고리즘영향평가를 의무화하였다.

특히 인권영향평가는 인공지능의 인권 위험을 식별하고 방지하는 데 유용한 도구로 주목받으며 도입되고 있다. 인권영향평가 절차는 대체로 계획 및 정보 수집 단계, 영향 식별 및 분석 단계, 영향 방지 및 완화 단계, 보고 및 점검 단계 등으로 구성된다고 볼 수 있으며, 이해관계자 참여와 인권에 미치는 위험과 그 심각도를 식별하는 것이 매우 중요하다.

평가 시점은 대체로 배치 이전에 실시하도록 하되 지속적인 모니터링과 조치 의무가 강조된다. 캐나다 알고리즘영향평가의 경우 프로젝트 설계 단계 초기에 우선 실시하고, 시스템의 생산 전에도 두 번째로 실시하여 요구사항이 구축된 시스템에 반영되었는지 확인하도록 하였다. 유럽연합 인공지능법(안)은 고위험 인공지능의 출시 전 위험평가 및 조

치, 적합성 평가와 CE 인증마크 표준 준수를 요구하였으며, 출시 후에도 모니터링과 품질관리를 요구하고 있다.

공개된 영향평가도구의 유형을 살펴보면, 캐나다 사례에서처럼 대상 기관이 자체적으로 평가하는 체크리스트 형식을 취하는 경우가 있고, 네덜란드 기본권 알고리즘영향평가에서처럼 인권 전문가가 내외부 이해관계자의 참여 속에서 판단해 가는 질의 형식을 취하는 경우가 있었으며, 나아가 덴마크 사례에서처럼 다른 평가도구들과 상호 보완하거나 통합적으로 최적화할 수 있도록 평가도구를 개방적으로 운영하는 경우도 있었다. 특히 전문성이 요구되는 인공지능 인권영향평가의 경우 해당 인공지능을 잘 알고 있는 사업담당자가 평가 절차에 참여하여 의견을 개진할 수 있겠지만, 인권에 미치는 영향을 평가하는 주체는 인권 전문가로서 평가 대상과 독립성을 보장하는 것이 바람직할 것이다. 또한 인권영향평가는 개인정보보호 영향평가 등 다른 평가와 보완적일 수는 있으나 인권에 미치는 부정적 영향에 대해서는 고유하게 평가할 필요가 있고, 기관이나 기업에서 구축한 전체적인 인권실사 및 인권영향평가 추진체계가 있다면 인공지능 인권영향평가 역시 그 절차 안에 통합되어 운영되는 것이 바람직할 것이다.

한편, 인공지능의 특성상 이에 대한 인권영향평가는 데이터에 대한 위험과 알고리즘에 대한 위험을 나누어 식별하고 완화 조치를 모색할 필요가 있다. 네덜란드 기본권 알고리즘영향평가는 데이터를 평가할 때 1) 관련 알고리즘 유형, 2) 데이터 원천 및 품질, 3) 데이터 편향성/가설, 4) 보안 및 보관에 대하여 평가하도록 하였고, 미국 알고리즘 책무성법(안)은 알고리즘 평가에서 데이터 최소화 및 보관기간 등 데이터의 개인정보 위험 및 개인정보보호 강화 조치를 살펴보고, 데이터 취득 시기 및 방법, 라이선스가 부여되었는지 여부 등을 살펴보도록 하였다.

알고리즘과 관련하여서는 네덜란드 기본권 알고리즘영향평가의 경우 알고리즘을 평가할 때 1) 알고리즘 유형(지도학습 여부), 2) 소유권 및 통제권, 3) 알고리즘 정확성, 4) 투명성 및 설명가능성을 평가하도록 하였다. 더불어 알고리즘에 기반한 의사결정과 인간의 역할을 평가하도록 하였다. 미국 알고리즘 책무성법(안)은 프로세스의 테스트 및 배치 조건에서 해당 프로세스의 성능을 검토하도록 하고, 모든 차별적인 수행에 대하여 검토하도록 하였다. 또한 소비자 권리 보호를 위하여 해당 프로세스가 사용될 것이라고 명확한 고지가 이루어졌는지, 이러한 사용에서 제외(opt-out)될 수 있는 방법이 있는지, 해당 프

로세스의 투명성과 설명 가능성은 물론, 소비자가 결정에 대해 이의제기·정정·재심을 청구하거나 해당 프로세스에서 제외될 수 있는 정도를 평가하도록 하였다.

한편, 인권영향평가에서는 식별된 위험의 방지 및 완화가 매우 중요하다. 유럽연합 인공지능법(안)은 고위험 인공지능의 위험은 적합한 설계와 개발을 통해 최대한 제거 또는 완화하여야 하며, 제거할 수 없는 잔여위험의 경우 허용가능 수준에 그쳐야 하고 사용자에게 통지되어야 한다고 규정하였다. 미국 알고리즘 책무성법(안) 역시 소비자에게 미치는 중대한 부정적 영향 가능성을 식별하고 식별된 중대한 부정적 영향 가능성을 제거하거나 합리적으로 완화하기 위해 취한 조치와 교육 방법을 문서화하도록 하였다.

마지막으로 인권영향평가는 그 결과를 이해관계자들에 대하여 공개하고 소통하는 것이 중요하다. 네덜란드 기본권 알고리즘영향평가의 경우에도 알고리즘의 사용에 대하여 소통할 계획을 갖추도록 하고, 다양한 집단이 알고리즘 결과물을 이해할 수 있도록 시각화할 것을 제시하였다. 미국 알고리즘 책무성법(안)은 영향평가에 대한 요약 보고서를 공개하도록 하였다.

한편, 최근의 공공기관과 기업들이 시행하는 인권영향평가의 경우 다양한 평가도구들을 최적화하여 사용하고 있다. 기관과 기업의 인공지능 사업에 대한 인권영향평가의 경우에도 대상 기관 및 기업들의 인권실사 및 인권영향평가 체계 속에서 인공지능 인권영향평가에 부합하는 기준 및 도구들을 결합하고 최적화하여 사용할 수 있을 것이다.

다만 최근 세계 각국은 자발적인 보고 의무를 중심으로 한 현행 인권실사제도가 효과적이지 못하고 관행을 변화시키지 못한다는 비판적 문제의식 속에 인권실사 실시 의무를 법적으로 부과하기 위해 노력하고 있다.[119] 우리나라에서도 인권실사의 실질화 또는 의무화에 대한 논의가 활발하게 이루어지고 있는 만큼, 인공지능 인권영향평가 역시 인권실사의 제도화 흐름과 조응할 필요가 있다.

이때 국가인권위원회는 인권경영의 확산을 위하여 인권영향평가의 기준과 도구를 제시하는 역할을 해 왔다. 국가인권위원회 진정 절차 또한 기관 및 기업의 사업활동과 관련된 인권침해 피해자가 이용할 수 있는 원상회복 및 고충처리 절차의 하나이다. 국가인권위원회는 인권침해행위 및 차별행위를 조사하여 구제에 나설 수 있으며, 기관과 기업은 국가인권위원회나 여타 감독기구 요구 시 인공지능 인권영향평가에 관한 자료를 성실

119) 김동현(2022), 120면.

하게 제출해야 할 것이다. 국가인권위원회는 권고 또는 의견의 표명으로 인공지능 인권영향평가의 법령・제도・정책・관행과 관련한 개선에 기여할 수도 있다.

인권영향평가를 비롯하여 인공지능 영향평가와 관련된 국내외 사례들은 아직 도입 초창기에 머물러 있고, 그 성과 또한 개별적이고 선언적인 모습을 보이고 있는 것이 사실이다. 그러나 향후 인공지능에 대한 다양한 평가절차 및 도구가 보다 종합적이고 모범적인 사례를 형성하고 국제기구와 국가인권위원회 등의 노력으로 실행력을 갖춘다면, 인공지능 인권영향평가 역시 인공지능이 인권에 미치는 부정적 영향과 위험을 식별, 완화, 방지하는 실질적인 규범력을 갖추어갈 수 있을 것이다.

제4장 심층면접조사

제1절 심층면접조사 개요

1. 조사 대상

본 연구는 '인공지능 인권영향평가' 수행을 위한 현실적이고 올바른 절차와 기준을 수립하는 것을 목적으로 하고 있다. 본 연구의 일환으로 <인공지능 인권영향평가(초안)>에 대해 관련한 전문가와 당사자의 다양한 의견을 수렴하고자 하였다. 조사는 기술, 정책, 공공분야 전문가와 당사자(관련단체)를 대상으로 서면 의견을 받는 방식으로 수행하였다. 28명의 전문가와 당사자에게 발송되었으며, 총 24명의 서면 의견지가 회수되었다. 기간은 2022년 11월 1일부터 11월 3일까지 배포를 시작하여 11월 18일에 마감하였다. 조사 대상은 다음과 같이 구성되었다.

[표 12] 심층면접조사 대상

구분	분야	인원(명)	구분	분야	인원(명)
당사자 및 관련단체	여성단체	1	기술전문가	기술	2
	빈곤단체	1	정책전문가	법률	2
	장애인당사자	1		기업	3
	교육단체	1		학술연구	6
	이주단체	1	공공기관	공공	3
	지역인권단체	3			
조사기간	2022. 11. 1. - 11. 3. 배포, 11. 18 취합				
조사방법	인공지능 인권영향평가도구(안) (서면 의견조사지)				

2. 조사문항 설계

조사문항은 인공지능 인권영향평가도구(안)의 절차 및 각 점검항목에 대해 전문가 및 당사자가 자유롭게 의견을 제시할 수 있도록 하였다. 그 외에 보완되어야 할 점검항목이 있는지 및 전반적인 추가의견을 질의하였다. 각 점검항목별로 질의 맥락에 대한 설명 및 국내외 사례에 대한 참고 사항이 포함되어 있지만, 여기서는 생략했음을 밝힌다. 질의 맥락에 대한 설명 및 참고 사항은 제5장의 인공지능 인권영향평가도구(안) 최종본에서 확인할 수 있다.

인공지능 인권영향평가도구(초안)

인공지능 인권영향평가의 절차

본 인공지능 인권영향평가 제도의 주요 형식 및 절차는 다음과 같다.

가. 인공지능 인권영향평가의 대상 : 고위험 인공지능

인권 침해의 위험성이 높은 인공지능으로서 적어도 아래 각 항목에 해당하는 인공지능은 일응 '고위험인공지능'에 해당한다고 보아 인공지능 인권영향평가의 대상으로 삼아야 할 것이다. 이는 EU AI 법안에서 정한 고위험 인공지능 분류를 따르면서 군대 및 정보기관이 사용하는 인공지능의 경우도 추가한 형태이다. 다만, 이는 현 단계에서 예측되는 잠재적 위험을 기준으로 한 것으로 확정적인 기준이 될 수 없고 하나의 예시로서 제안되는 것이므로, 향후 위험성이 커보이는 분야가 새롭게 확인되거나 드러나는 경우 항목이 추가될 수 있고, 반대의 경우 제외될 수 있다.

가. 인공지능이 항공, 자동차, 철도, 기계, 장난감의 안전, 승강기, 무선 장비 및 의료 기기 등 제품 자체 또는 제품의 안전요소인 경우
나. 실시간 또는 사후적으로 사람의 생체정보를 활용하여 신원확인을 수행하는 경우
다. 도로, 교통, 물, 가스, 전기 공급 등 중요 인프라 관리·운영의 안전 구성요소로 사용되거나 소방관, 응급의료 등의 긴급 응급 대응 서비스 파견 또는 우선 순위를

설정하는데 사용되는 경우
라. 적합한 교육·직업훈련기관 선정 및 지원 결정, 교육생 및 훈련생 평가에 사용되는 경우
마. 결원공고, 지원서 선별, 후보자 평가, 승진 결정, 작업 할당, 업무 성과 모니터링 등
바. 공공 지원 혜택과 서비스(수당, 감면, 유예, 환원 등) 자격 및 수혜 적격성을 평가하기 위하여 공공기관에서 사용하는 경우
사. 경찰 등이 법 집행 목적으로 인공지능을 이용하는 경우
아. 거짓말 탐지기 및 유사한 도구로 사용, 사람의 감정 상태 감지, 입국시 보안위험, 건강 위험 및 불법체류 가능성 평가, 망명, 비자, 영주권 신청의 적격성 평가, 입국 및 여행 서류의 진본 여부 검증 등에 사용되는 경우
자. 사실 및 법률 조사, 법률 해석 및 적용 지원 등 법정에서 판사를 지원하는 데 사용되는 경우
차. 군 또는 국가 정보기관이 인공지능을 사용하는 경우

법률에서 금지하는 인권 침해 또는 차별 대우를 목적으로 하거나 법률에서 금지하는 개인정보의 처리를 목적으로 하는 '금지대상 인공지능'은 위험의 완화 내지 제거가 불가능하므로 인권영향평가의 대상에서 제외된다.

나. 인공지능 인권영향평가 시기

공공과 민간을 막론하고 고위험 인공지능을 개발하거나, 고위험인공지능을 사업 또는 정책의 기반기술로 도입하기 이전에 인권영향평가를 수행하되, 사전영향평가에만 한정하지 않고, 정기적, 사후적 평가를 통한 지속적인 관리와 모니터링을 전제한다. 고위험인공지능에 비하여 위험의 정도가 덜한 인공지능의 경우 국가인권위원회의 직권 지정 또는 자발적인 요구에 의해 인권영향평가를 수행할 수 있다.

다. 인공지능 인권영향평가 수행 주체

구체적인 영향평가의 수행은 독립성과 인권 분야에 대한 전문성을 갖춘 제3의 기관이 수행하도록 한다. 영향평가에 따른 결과서는 국가인권위원회에 제출되며, 국가인원위원회는 영향평가결과를 검토한 후 미흡한 점에 대한 개선을 권고하거나, 위험성에 대한 완화 조치 또는 제거 조치가 불가능하다고 판단하는 경우 개발 또는 활용의 중단을 권고하는 등 의견을 제시할 수 있도록 한다.

라. 인공지능 인권영향평가의 절차

본 연구에서는 이상의 검토내용을 토대로 인공지능 인권영향평가의 이행단계를 아래와 같이 4단계로 범주화하였다.

1단계 : 계획, 범위설정, 조사
2단계 : 영향 분석 및 평가
3단계 : 방지, 완화, 구제
4단계 : 공개 및 점검
공통 : 이해관계자의 참여

첫 번째 계획, 범위설정, 조사 단계에서는 영향평가를 수행할 주체를 구성하고 관련한 기본적인 문헌을 조사하며, 평가 대상이 되는 인공지능 관련 사업 또는 정책을 선정하고 관련 이해당사자를 식별한다. 대상 기술 또는 사업과 관련한 기본 사실을 기입하면서 사업 계획 또는 진행 정도가 충분히 성숙되었는지 여부 등도 확인한다.

영향 분석 및 평가 단계에서는 데이터, 알고리즘, 심각도의 각 분류 항목에 따른 체크리스트에 따라 그 영향의 정도를 평가한다. 이 단계는 인공지능 시스템에 의해 부정적인 영향을 받을 가능성이 있는 관련 권리를 식별하고 이에 대하여 영향의 정도까지 분석하는 과정을 포함한다. 방지 및 완화, 구제 단계에서는 개선 조치와 완화 조치가 이행되었는지를 확인한다.

공개 및 점검 단계에서는 완화조치들이 부정적 영향을 완화하였는지 지속적으로 평가하고 점검하는 단계이다. 투명성과 설명가능성을 충족하는지, 영향평가 결과를 공개하였는지 여부 등을 확인한다.

국내외 많은 영향평가에서 권고하고 있듯이, 이해관계자의 참여는 특정한 단계가 아니라 모든 단계에서 고려되거나 이행해야 할 사항이다. 예를 들어, 1단계에서는 인공지능 시스템의 성격이나 위험성에 따라 관련 이해관계자들을 식별하고 해당 이해관계자들과의 협의를 통해 정보를 수집하고 침해 가능성이 있는 인권을 식별하는 작업이 이루어져야 한다. 3단계(방지 및 완화)에서도 인권 침해를 방지, 완화하고 피해자의 권리를 구제하기 위한 여러 정책 제안에 대한 이해관계자의 의견을 수렴할 필요가 있다.

아래에서는 인권영향평가 과정에서 점검해야 할 항목을 체크리스트 방식으로 제시한다. 영향평가 수행자는 각 점검항목에 대해 “예 / 보완 필요 / 아니오 / 정보 없음 / 해당 없음” 중 하나를 선택할 수 있다. 그러나 인권영향평가는 정답을 맞추는 것이라기 보다는 이해관계자 사이의 소통과 토론, 그리고 서로의 역량을 강화하는 과정이 되어야 한다. 또한 인권영향평가 결과의 공개가 권고된다는 점에서, 평가 대상이 되는 인공지능 시스템이나 프로젝트에 대해 관련 이해관계자들이 이해하기 쉽도록 정보를 제공할 필요가 있다. 이에 모든 항목에 대해서 체크리스트 방식의 표기와 함께, 관련된 구체적인 정보를 서술식으로 설명하는 공간을 제공하고 있다. 이를 활용하여 기술에 전문성이 없는 이해관계자라도 이해할 수 있도록 가능한 쉽고 상세한 설명을 제공할 것을 권장한다.

1단계 : 계획, 범위설정, 조사

가. 인권영향평가 계획

Q1-1-1. 조직 내에 인공지능 인권영향평가 수행 관련 정책을 두고 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q1-1-2. 인권영향평가를 수행하는 팀의 역할과 목표, 인권영향평가의 결과를 누구에게 어떻게 보고하는지가 명확하게 규정되어 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q1-1-3. 인권영향평가를 내실있게 수행할 수 있도록 평가 대상이 되는 인공지능 시스템의 개발 혹은 활용과 관련한 부서 및 담당자와의 협력과 자료에 대한 접근 권한이 보장되어 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q1-1-4. 인권영향평가를 수행하는 책임자는 누구입니까

책임자의 성명과 소속 :

Q1-1-5. 인권영향평가의 대상이 되는 인공지능 시스템 혹은 프로젝트는 무엇입니까

인공지능 시스템 혹은 프로젝트명 :

Q1-1-6. 인공지능 시스템은 어떠한 문제를 해결하기 위한 것입니까, 인공지능 시스템이 달성하고자 하는 목적 및 의도된 용도는 무엇입니까

인공지능 시스템의 목적 :

나. 범위 설정

Q1-2-1. 인공지능 시스템의 도입 및 활용, 혹은 인공지능 시스템을 통한 의사결정과 관련된 법적인 근거가 있습니까. 있다면 그 법적 근거는 무엇입니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q1-2-2. 인권영향평가 외에 인공지능 시스템과 관련된 다른 기준이나 인공지능 시스템이 적용될 분야의 다른 규범을 검토하였습니까

Q1-2-2-1. 해당 조직이 국가인권위원회의 <인공지능 개발과 활용에 관한 인권 가이드라인>을 검토하였는지 확인하였습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q1-2-2-2. 개발자가 개인정보보호위원회의 <AI 개인정보보호 자율점검표>에 따라 점검하였는지 확인하였습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q1-2-2-3. 개발자가 과학기술정보통신부 <신뢰할 수 있는 인공지능 개발 안내서(안)>에 따라 점검하였는지 확인하였습니까?

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q1-2-2-4. (금융기관의 경우) 금융위원회 <금융분야 AI 개발·활용 안내서>를 점검하였는지 확인하였습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q1-2-2. 인권영향평가 외에 인공지능 시스템과 관련된 다른 기준이나 인공지능 시스템이 적용될 분야의 다른 규범을 검토하였습니까

Q1-2-2-1. 해당 조직이 국가인권위원회의 <인공지능 개발과 활용에 관한 인권 가이드라인>을 검토하였는지 확인하였습니까

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q1-2-2-2. 개발자가 개인정보보호위원회의 <AI 개인정보보호 자율점검표>에 따라 점검하였는지 확인하였습니까

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q1-2-2-3. 개발자가 과학기술정보통신부 <신뢰할 수 있는 인공지능 개발 안내서(안)>에 따라 점검하였는지 확인하였습니까?

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q1-2-2-4. (금융기관의 경우) 금융위원회 <금융분야 AI 개발·활용 안내서>를 점검하였는지 확인하였습니까

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q1-2-2-5. 해당 조직이 개인정보 영향평가를 수행하였는지 확인하였습니까. 수행하였다면 영향평가 결과보고서를 확보하고 있습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q1-2-2-6. 해당 조직이 성별 영향평가를 수행하였는지 확인하였습니까. 수행하였다면 영향평가 결과보고서를 확보하고 있습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q1-2-2-7. 인공지능 시스템이 활용될 분야에서 적용되는 다른 규범이나 기준이 있습니까. 있다면 그것은 무엇입니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q1-2-3. 초기 영향평가 과정에서, 침해될 우려가 있는 것으로 파악된 인권은

무엇입니까. 침해될 우려가 있는 인권을 모두 적어주세요.

답변 :

Q1-2-4. 인공지능 시스템과 관련된 이해관계자가 누구인지 파악하였습니까. 가능한 구체적으로 적어주세요.

Q1-2-4-1. 인공지능 시스템의 기획자 및 개발자는 누구입니까

Q1-2-4-2. 인공지능 시스템의 유지, 보수를 책임지는 주체는 누구입니까

Q1-2-4-3. 인공지능 시스템의 의도된 사용자는 누구입니까

Q1-2-4-4. 인공지능 시스템의 사용으로 영향을 받는 사람이나 집단은 누구입니까

Q1-2-4-5. 인공지능 시스템의 사용으로 영향을 받는 개인이나 집단에 아동, 노인, 장애인, 여성, 외국인, 성소수자, 저학력자, 경제적 약자, 낙후지역 등 취약하거나 소외된 집단이 포함되어 있습니까

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q1-2-4-6. 조직 내부에서 이 인공지능 시스템의 개발 및 운영에 관련된 기타 이해관계자는 누구입니까 (정책부서, 데이터 거버넌스 부서, 영업부서 등)

Q1-2-4-7. 조직 외부에서 이 인공지능 시스템의 개발 및 운영에 관련된 기타 이해관계자는 누구입니까 (위탁업체, 감독기구, 전문가 집단 등)

다. 조사

Q1-3-1. 인공지능 시스템 개발을 위해 활용된 데이터셋과 알고리즘은 무엇입니까? 해당 데이터셋이나 알고리즘의 특성, 목적 및 용도, 장점 및 단점에 대한 정보를 문서로 확보하고 있습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q1-3-2. 인공지능 시스템이 도입, 활용될 분야 혹은 지역적인 특성 및 맥락과 관련된, 인권에 영향을 미칠 수 있는 요소에 대한 자료를 확보하고 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q1-3-3. 앞서 파악한, 인공지능 시스템의 이해관계자로부터 해당 시스템이 인권에 미칠 영향에 대한 의견을 수렴하거나 협의하고 이를 문서화 하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q1-3-4. 이해관계자 의견을 수렴하거나 협의할 때 다음과 같은 내용을 포함합니까
- 협의한 이해관계자의 성명, 소속, 연락처
- 협의한 일자
- 인공지능 시스템에 대해 이해관계자에게 제공한 자료
- 인공지능 시스템에 대한 이해관계자의 의견

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q1-3-5. 인공지능 시스템의 활용으로부터 영향을 받는 이해관계자, 특히 취약하거나 소외된 집단의 협의를 포함하고 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q1-3-6. 관련 자료를 수집하거나 이해관계자의 의견을 수렴할 때, 필요한 경우 자료의 기밀성을 유지하고 이해관계자의 개인정보를 보호할 수 있는 조치를 취하고 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

2단계 : 영향 분석 및 평가

가. 인공지능 기술과 관련된 영향 분석 및 평가

(1) 개인정보보호

Q2-1-1. 인공지능 시스템이 개인정보보호위원회 <AI 개인정보보호 자율점검표>의

모든 의무/권장 조항을 준수하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

(2) 데이터

Q2-1-2. 학습, 검증, 테스트 등 인공지능의 개발 과정에 사용되는 데이터셋은 그 출처, 구조와 유형, 사전 처리 과정이 문서화되어 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q2-1-3. 데이터셋의 정확성, 완전성, 최신성을 확인하였습니까. 이를 검토하기 위해 사용한 방법은 무엇입니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q2-1-4. 데이터셋이 인공지능 시스템이 사용될 맥락에 적합하도록 인구집단별 다양성과 대표성을 갖추었는지 확인하였습니까. 이를 검토하기 위해 사용한 방법은 무엇입니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q2-1-5. 데이터셋이 민감한 개인정보를 포함하고 있습니까. 대리 변수를 통해 민감한 개인정보의 추정이 가능한지 여부를 검토하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

(3) 알고리즘의 성능과 신뢰성

Q2-1-6. 인공지능 시스템에 사용된 알고리즘이 목적 달성에 가장 적합한 이유는 무엇입니까? 채택된 알고리즘 외에 다른 알고리즘 혹은 인공지능 시스템 외의 다른 대안에 대한 검토가 있었습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q2-1-7. 인공지능 시스템이 의도한대로 작동하는지 성능을 측정하기 위한 지표와 방법을 갖고 있습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q2-1-8. 인공지능 시스템의 정확도와 오류율은 어떻게 측정합니까. 정확도와 오류율의 수준은 의도한 목적에 적합한 정도로 설정되었습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

(4) 차별금지

Q2-1-9. 인공지능 시스템이 활용 과정에서 성별, 종교, 장애, 나이, 출신 지역, 신체조건, 피부색, 성적 지향, 사회적 신분 등 개인과 집단의 특성에 따라 특정 집단에 대한 차별을 야기하거나 혹은 기존의 차별을 악화할 가능성이 있는지 검토하였는가.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q2-1-10. 인공지능 시스템 개발 과정에서 알고리즘에 의한 구조적 차별을 사전에 방지하기 위하여 개발팀 구성원의 다양성 확보, 개발자에 대한 교육, 조직 내 인공지능 윤리 정책 수립 등의 대책을 마련하고 있는가.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

(5) 설명가능성과 투명성

Q2-1-11. 인공지능 시스템이 특정한 결정(출력)을 내리는데 사용된 데이터 혹은 요소를 추적할 수 있습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q2-1-12. 인공지능 시스템이 특정한 결정(출력)을 내린 이유나 근거에 대해 사용자 혹은 영향을 받는 이해관계자에게 설명할 수 있습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음

비고 ( )

Q2-1-13. 인공지능 시스템의 작동이나 특정한 결정의 근거에 대해 기술 전문가가 아닌 이해관계자가 충분히 이해할 수 있는 방식으로 설명할 수 있습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q2-1-14. 인공지능 시스템의 성능(정확도, 오류율 등), 어떤 결정을 내리는데 사용되는 매개변수 및 가중치, 적절한 사용법, 장점과 한계 등에 대해 사용자가 이해할 수 있는 방식으로 충분한 정보를 제공하고 있습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q2-1-15. 인공지능 시스템의 소스코드가 공개되거나 이를 요구하는 이해관계자에게 제공될 수 있습니까. 소스코드가 제공될 수 있다면, 누구에게 어떤 조건으로 제공됩니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

(6) 자동화 정도와 인간의 개입

Q2-1-16. 알고리즘 시스템의 사용자 혹은 알고리즘과 상호작용하는 주체가 자신이 상호 작용하는 것이 인공지능 시스템이라는 사실 혹은 자신에게 내려진 결정이 인공지능 시스템에 의한 것이라는 점을 명확하게 인식할 수 있도록 표시하고 있습니까

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q2-1-17. 인공지능 시스템이 영향을 받는 당사자가 인지하지 못하는 방식으로 작동할 수 있습니까. 영향을 받는 당사자가 인지하지 못하는 방식으로 인공지능 시스템이 작동하지 않도록 당사자에게 적절하게 알릴 수 있는 조치를 취하고 있습니까

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q2-1-18. 인공지능 시스템의 결과물에 기반한 의사결정에서 인간은 역할과 인간이

재량권을 갖고 개입할 수 있는 범위와 절차가 정의되어 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q2-1-19. 인공지능 시스템이 내린 의사결정에 의해 영향을 받는 당사자가 이를 거부하거나 결정에 이의를 제기할 수 있는 수단이 마련되어 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q2-1-20. 인공지능 시스템이 의도한대로 작동하지 않을 경우 사용자는 언제든지 시스템을 정지시킬 수 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

(7) 보안

Q2-1-21. 인공지능 시스템의 특성과 활용되는 분야 등을 고려했을 때, 인공지능 시스템 보안에 대한 가능한 위협이 무엇이고 보안이 침해되었을 경우 발생할 수 있는 결과 혹은 해악은 무엇입니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q2-1-22. 인공지능 시스템의 학습 및 테스트에 활용되는 데이터셋에 대해 충분한 안전조치가 적용되었습니까. 데이터 오염 등 데이터에 대한 다양한 유형의 공격에 대한 대응이 고려되었습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q2-1-23. 인공지능 시스템의 전체 수명주기 동안 발생할 수 있는 잠재적인 공격에 대비하여 무결성, 견고성, 전반적인 보안을 보장하기 위한 조치를 취했습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

(8) 접근성

Q2-1-24. 인공지능 시스템이 특별한 필요나 장애가 있는 사람들도 사용할 수 있도록 인터페이스가 보편적 설계(유니버설 디자인) 원칙에 따라 설계되었습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

(9) 알고리즘 라이선스

Q2-1-25. 인공지능 시스템의 알고리즘이 외부에서 개발된 경우, 알고리즘의 소유권 및 관리 권한에 대해 명확한 합의가 이루어져 있습니까. 해당 조직이 원할 경우 알고리즘을 적절하게 변경할 수 있는 권한이 부여되어 있습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

나. 인권에 미치는 영향 및 심각도

(1) 영향을 받는 인권

Q2-2-1. 인공지능 시스템이 도입, 활용될 경우 시민들의 인권에 미칠 수 있는 부정적인 영향의 위험은 무엇입니까. 누구의 인권이 어떤 방식으로 침해될 수 있습니까. 인공지능 시스템이 여러 인권(들)에 부정적 영향을 미칠 수 있는지에 대해 모두 파악이 되었습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q2-2-2. 인공지능 시스템이 오류로 인하여 의도하지 않은 방식으로 작동할 경우 나타날 수 있는 부정적인 결과에 대해 검토한 바 있습니까. 의도하지 않은 방식으로 작동할 경우 침해되는 인권은 무엇입니까

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

Q2-2-3. 인공지능 시스템이 의도적으로 악용될 가능성이 있습니까. 어떠한 방식으로 오용될 수 있는지에 대해 검토하였습니까. 악용될 경우 나타날 수 있는 부정적인 결과, 혹은 침해되는 인권은 무엇입니까

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
비고 ( )

(2) 인권에 미치는 영향의 심각도

Q2-2-4. 인공지능 시스템이 인권에 미치는 부정적 영향의 범위에 대해 검토하였습니까. 부정적 영향이 전체 인구 혹은 특정 집단에 미치는 범위가 어떠합니까. (부정적 영향을 받을 수 있는 인권이 여러 개인 경우 각각에 대해서. 아래 질의에 대해서도 동일함)

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q2-2-5. 인공지능 시스템이 인간의 생명, 인권, 기본적 삶 등에 미치는 부정적 영향의 규모 혹은 크기에 대해 검토하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q2-2-6. 인공지능 시스템이 인권에 미치는 부정적 영향이 사후에 구제나 회복이 가능한지 여부에 대해서 검토하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q2-2-7. 인공지능 시스템이 인권에 미치는 부정적 영향의 범위, 규모, 회복불가능성 등이 모두 높게 평가될 경우 인공지능 시스템의 도입을 철회할 수 있는 절차가 마련되어 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q2-2-8. 인공지능 시스템이 인권에 미치는 부정적 영향이 여러 개일 경우 그 심각성에 따라 우선 순위를 부여할 수 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

3단계 : 방지, 완화, 구제

가. 방지

Q3-1-1. 데이터에 대한 개선, 알고리즘의 수정, 시스템 설계 변경 등 2단계(영향 분석 및 평가)에서 파악된 중대한 인권 위험을 방지하기위한 조치를 취하고 문서화하였습니까. 그렇다면 어떤 조치를 취했습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

나. 완화

Q3-2-1. 2단계(영향 분석 및 평가)에서 파악된 중대한 인권 위험을 완화하기 위한 조치를 취하고 문서화하였습니까. 그렇다면 어떤 조치를 취했습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q3-2-2. 인공지능 시스템의 잔존하는 위험성에 대해 사용자에게 충분한 정보를 제공하고 올바른 작동 방법에 대해 적절한 교육을 제공하고 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q3-2-3. 인공지능 시스템의 인권 침해 위험이 클 수 있는 특정한 사용을 허용하지 않도록 이용약관이나 여타의 집행체계에서 금지하는 절차를 취하고 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q3-2-4. 인공지능 시스템의 인권 침해 위험을 조속히 탐지할 수 있는 절차나 감독 메커니즘을 취하고 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q3-2-5. 2단계(영향 분석 및 평가)에서 파악된 중대한 인권 위험이 완화되지 않고 남아있을 경우 그 이유를 문서화하고 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q3-2-6. 중대한 위험에 대한 방지 및 완화 조치를 취하기 힘들거나, 이러한 조치를 취해도 여전히 중대한 위험이 남아있을 경우 인공지능 시스템의 개발 및 활용을 중단할 계획을 갖고 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

다. 구제

Q3-3-1. 인공지능 시스템의 결정에 의해 영향을 받는 사람이 인공지능 시스템의 결정에 이의를 제기하거나 침해된 권리의 구제를 요구할 수 있는 절차가 마련되어 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q3-3-2. 인공지능 시스템에 의해 영향을 받는 사람들에게 인공지능 시스템의 사용을 거부할 수 있는 선택권(옵트아웃 권리)을 제공하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q3-3-3. 인공지능 시스템의 결정에 대한 이의제기나 권리구제 요구가 정당할 경우, 문제의 의사결정을 번복하거나 권리를 복구하거나 손해배상을 할 수 있는 절차가 마련되어 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

라. 이해관계자와의 의견수렴 및 협의

Q3-4-1. 인공지능 시스템의 인권 위험을 방지, 완화하고 인권을 침해받은 사람의 권리를 구제하기 위한 조치에 대해서 관련 이해관계자 의견을 수렴하거나 협의를 진행하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

4단계 : 공개 및 점검

가. 인공지능 시스템의 주요 요소의 공개

Q4-1-1. 인공지능 시스템이 사용하는 데이터와 알고리즘의 주요 요소를 일반에 공개하고 이해할 수 있는 방식으로 쉽게 설명하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

나. 인권영향평가 결과 공개

Q4-2-1. 인권영향평가 보고서 전체 혹은 주요 내용을 일반에 공개합니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q4-2-2. 인권영향평가 보고서를 감독기구인 국가인권위원회에 제공하고, 효과와 한계에 대해 협의하는 절차가 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

다. 인공지능 시스템에 대한 모니터링

Q4-3-1. 인공지능 시스템이 도입되거나 운영이 시작된 후에 그 성능과 인권에 미치는 부정적 영향, 완화 조치 및 구제 정책의 효과성을 확인하기 위해, 인공지능 시스템의 수행을 모니터링하고 기록에 남길 수 있는 메커니즘을 수립하였습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q4-3-2. 인공지능 시스템이 의도한대로 작동하지 않거나 인권에 미치는 부정적인 영향이 확인되었을 때, 관련된 책임을 명확히 하고 인공지능 시스템을 개선하며 부정적 영향을 완화하기 위해 필요한 절차를 수립하였습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

라. 인권영향평가에 대한 점검

Q4-4-1. 인권영향평가 수행의 효과와 한계를 점검할 수 있는 절차를 마련하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

마. 인권영향평가의 반복적 수행

Q4-5-1. 인공지능 시스템에 대해 정기적으로 (예를 들어 1년) 인권영향평가를 수행하는 절차를 마련하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

Q4-5-2. 인공지능 시스템의 핵심적인 기능이 변경되거나 환경적 요인 혹은 적용 범위가 변경되었을 경우, 인권영향평가를 다시 수행하는 절차를 마련하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
비고 ( )

제2절 개별서면조사 결과

1. 인공지능 인권영향평가 절차(개요)에 대한 의견

귀하께서는 인공지능 인권영향평가도구(안)의 영향평가 절차에 대해 어떻게 생각하십니까? 보완 및 개선을 위한 내용을 작성해주십시오.

가. 절차(개요)에 대한 전반적인 의견

많은 응답자들이 인공지능 인권영향평가 도입의 필요성에 대해서는 공감하였다. <인공지능 개발과 활용에 관한 인권 가이드라인>의 취지와 국제적 기준들이 '인공지능 인권영향평가도구(안)' 에 반영되었다는 의견을 비롯하여, 문항이 구체적이고 설명도 상세하여 도움이 될 것 같다는 의견, 전반적으로 체계적이고 꼼꼼하게 작성되었다는 의견, 구체적인 질문을 순서에 따라 제시하여 주요 사항들이 잘 정리되어 있다는 의견 등이 있었다.

반면, 일부 기업 분야 응답자들은 인권영향평가를 인공지능의 인권침해 위험을 해소하는 하나의 방법으로 인식하기보다는 기업에 대한 또 하나의 불필요한 규제로 받아들이는 경향을 보였다. 또한 인권영향평가 과정에서 기업의 영업비밀이 침해될 수 있다는 점에 대해 큰 우려를 표명하였다. 일부 학술 분야 응답자들도 인권영향평가가 기업에게 규제로 작용할 수 있는 만큼, 공공부문에서 우선으로 시행할 필요가 있다는 의견을 제시하였다. 하지만 아직 인권영향평가의 의무 대상이 법제화되어 있는 단계는 아닌 만큼, 의무 대상에 대해서는 추후 심도 있는 논의가 필요할 것으로 보인다.

인공지능 인권영향평가도구(안)(이하, 도구(안)) 개요에 대한 응답자의 주요 의견은 다음과 같다.

1) 인공지능 인권영향평가의 대상에 대한 의견

응답자들은 이해관계자에 따라 서로 다른 의견을 제시하였다. 한 기업 분야 응답자는 영향평가의 대상이 되는 '고위험 인공지능' 을 예시적으로 제시하고 있는데 고위험을

판단할 수 있는 기준이나 방법론이 명확하지 않다며 위험 수준에 대한 별도의 분류 체계가 마련, 제시되어야 한다는 의견을 제시했다. 비슷한 맥락에서 한 학술 분야 응답자는 인권영향평가가 스타트업 등에게 하나의 진입장벽으로 작용하거나 진입 비용을 높이는 결과를 초래할 것이라는 우려를 제기했다.

"오히려 평가 자료의 공개를 예정하고 있고 향후 국가인권위원회의 후속 조치(의견 제시)까지 고려하고 있으시다면 정부 부처와 국책/공공기관의 AI 시스템 활용에 우선 적용하는 것으로 하고 이를 통해 경험을 축적하면서 영향평가도구와 절차를 더 정비한 뒤 적용 범위를 확장하는 방안에 대해서도 고려해 보시기 바랍니다. 특히 현재의 (안)에서 요구하고 있는 듯한 다양한 절차와 조치(문서화 포함)는 상당한 인적 조직과 비용을 전제하고 있어 본(안)이 적용될 관련 시장경쟁에서(현재의 (안)에 따르면 고위험 리스크 AI 관련시장) 스타트업 등에게 하나의 진입장벽으로 작용하거나 진입 비용을 높이는 결과를 초래할 염려도 있습니다." (학술연구자)

반면에, 인권단체 분야 응답자는 인공지능 인권영향평가 대상을 고위험 인공지능에 한정하고 있는데, 인권침해 소지가 크지만 고위험 인공지능에 해당하지 않을 경우 평가대상에서 배제될 가능성을 우려하였다. 한 학술연구자는 인권영향평가의 필요성에 대해서는 매우 긍정하는 입장이지만, 영향평가 대상이 법령상 의무화되어 있는 것도 아니어서 인권영향평가가 상징적인 권고 수준을 넘어서기 어렵다는 의견을 주면서 입법화의 필요성에 주목하였다. 법률 분야 응답자는 "제한된 위험의 인공지능도 경우에 따라 특정 개인이나 집단에 큰 영향을 줄 수 있으므로 평가수행의 요건에 '영향을 받는 이해관계자의 신청을 받아 국가인권위가 지정' 하는 경우를 추가" 할 것을 제안하였다. 대다수 응답자는 정부 부처와 국책/공공기관의 인공지능 시스템 활용에 인권영향평가를 적용해야 한다는 것에 이견이 없었다.

"특히 정부의 기금을 통해 개발되는 인공지능 시스템의 경우 반드시 인권영향평가를 수행할 수 있도록 하였으면 합니다. 공공의 기금이 인권을 침해하는 결과를 낳지 않도록 필수적으로 점검이 필요합니다." (인권단체 활동가)

2) 인공지능 인권영향평가 시기에 대한 의견

기술 분야 응답자는 평가가 완료된 이후에는 개선의 여지가 많지 않으므로, 실제 개발되기 이전인 인공지능 시스템 개념 도출 단계에서부터 가이드라인을 숙지하여 기획하는 것이 의무였으면 좋겠다는 의견과 인권전문가뿐만 아니라, 인공지능 기술전문가의 개입 필요성을 강조하였다.

3) 인공지능 인권영향평가 수행 주체에 대한 의견

여러 응답자들이 인권영향평가의 실행 주체가 누구인지 모호하다는 의견을 제시하였다. 절차에서는 '독립성과 인권 분야에 대한 전문성을 가진 제3의 기관'이 수행하는 것으로 표현되어 있으면서도 검토 항목에서는 '인권영향평가를 수행하는 팀의 역할과 목표를 규정'하도록 하여 마치 조직 내의 기관이 수행하는 것처럼 표현하여 혼란을 야기한 것으로 보인다. 인권영향평가는 상황에 따라 제3의 기관뿐만 아니라, 해당 조직 내의 독립적인 팀에서도 수행할 수 있는 것으로 정리하였으며, 이러한 혼란이 야기되지 않도록 질의 문구를 다듬을 필요가 있다.

"전체적으로 보았을 때, 먼저 인권영향평가의 실행 주체에 대한 고민이 필요해 보입니다. 기존의 인권영향평가를 보면, 조직의 내부 절차와 외부절차가 있을 수 있고, 내부 절차에서도 개발부서가 직접 수행하는 경우(가장 좁은 의미의 자율점검), 별도의 (독립적) 부서를 두는 경우 등 다양한 경우가 있습니다. 별도의 독립적 부서나 외부에서 수행할 경우 객관성, 독립성이 보장된다는 장점이 있지만, 반대로 전문성이 부족할 수 있고, 시간과 비용이 많이 소요된다는 단점도 있습니다. 그래서 어떤 경우에는 평가 대상의 중요성에 따라 여러 절차를 선택적으로 수행하기도 합니다. 이 도구를 이용하는 기관에서 평가 주체를 어떻게 설정하는 것이 좋을지에 대한 지침이 포함되는 게 좋겠다고 생각합니다." (학술연구자)

"구체적인 영향평가의 수행은 독립성과 인권분야에 대한 전문성을 갖춘 제3의 기관이 수행하도록 한다." 고 쓰여있는데, Q-1-2에서 인권영향평가를 수행하는 팀의 역할과

목표가 규정되어 있는지, ‘제3의 기관’과 ‘인권영향평가를 수행하는 팀’이 같은 것인지 다른 것인지 혼동이 있을 수 있겠습니다. Q-1-4에서 인권영향평가를 수행하는 책임자는 누구인지 묻는 것도 마찬가지입니다. 책임자는 “제3의 기관” 책임자를 말하는 것일까요?” (학술연구자)

4) 인공지능 인권영향평가 절차에 대한 의견

공공분야 응답자는 인공지능 인권영향평가의 절차를 단계별로 구성한 것은 인공지능이 인권에 대해 미치는 영향을 체계적으로 평가하고 해당 평가의 결과를 향후 활용하는 부분까지 고려하므로 의미가 있지만, 단계의 명칭이 명확하지 않아서 변경이 필요해 보인다는 의견을 제시해주었다.

“인공지능 인권영향평가의 절차를 단계별로 구성한 것은 매우 유의미한 것으로 보입니다. 다만, 현재 “1단계는 계획, 범위설정, 조사 2단계는 영향 분석 및 평가, 3단계는 방지, 완화, 구제 4단계는 공개 및 점검”으로 구성되어 있는 절차상 단계의 명칭이 어떠한 과정이 진행되는지가 명확하지 않습니다(특히, 1단계와 3단계). 1단계는 협의・준비 단계, 2단계는 분석・평가 단계, 3단계는 개선・완화 단계, 4단계는 공개・점검 단계” 정도로 명칭을 간결하게 하여 명칭만으로도 해당 단계에서 어떠한 절차가 진행되는지를 가늠할 수 있게 하면 어떨까 싶습니다.” (공공분야 응답자)

다른 공공분야 정책전문가는 인공지능 인권영향평가의 절차를 ‘순서도’로 정리하여 소개하면 좋을 것 같다는 의견을 주었다.

2. 인공지능 인권영향평가도구(안)의 주요 검토항목에 대한 의견

귀하께서는 인공지능 인권영향평가도구(안)의 주요 내용에 대해 어떻게 생각하십니까? 각 분야마다 추가 혹은 보완해야 할 질의 항목이 있다면 작성해주십시오.

가. ‘1단계’ 계획, 범위설정, 조사에 대한 의견

1) ‘계획’ 관련 검토항목에 대한 의견

앞서 수행 주체와 관련한 의견에서 언급했다시피, 1단계 ‘계획’ 관련 검토항목의 질의와 관련하여 인권영향평가 주체가 누구인지 혼동을 야기한다는 의견이 있었다. 또한, 이러한 맥락에서 어느 법률 분야 응답자는 협력과 자료 접근의 주체가 인권영향평가 수행팀이라는 점을 명확하게 명시할 필요가 있다는 의견을, 인권단체 및 기업 분야 응답자는 인권영향평가 결과를 누구에게 어떻게 보고하는지를 명확하게 규정될 필요가 있다는 의견을 제시하였다.

한편, 인권영향평가를 수행하기 위한 정책의 적절한 운용을 보장할 수 있는 예산 및 자원의 확보 여부에 대한 질의가 필요하다는 의견도 제시되었다.

“인권영향평가 수행 관련 정책의 적절한 운용을 보장하기 위한 인적, 물적 자원과 정책적 자원이 확보되어있는지에 대한 확인을 추가하는 것이 필요할 것으로 생각됩니다” (기업 분야 응답자).

그 외, 책임자가 관련 교육을 이수했는지, 인공지능 기술 및 인권에 대한 전문성을 갖추고 있는지를 질의해야 한다는 인권단체 활동가의 의견도 있었다.

2) ‘범위설정’ 관련 검토항목에 대한 의견

해당 인공지능 시스템과 관련된 법적 근거와 관련하여 여러 응답자로부터 법적인 근거를 질문하는 의도가 무엇인지 모호하다는 의견이 제시되었다. 질문의 취지는 해당 인공지능 시스템이 도입될 분야에 적용되는 법령이 있을 경우 인공지능 시스템에도 해당 법령이 적용될 것이므로 설계 시부터 이를 고려해야 한다는 것이었다. 예를 들어 공공기관에서 면접을 시행할 때 적용되는 요건이 있다면 인공지능 면접 도구에도 동일한 요건이 적용될 것이다. 이와 같은 질의의 취지가 명확하게 드러나도록 보완될 필요가 있다.

인권영향평가 외에 다른 기준이나 규범을 검토했느냐에 대한 질의에 대해서도 여러

비판적인 의견이 제시되었는데, 요컨대 다른 정부 기관이 관할하고 있는 규범이나 안내서와 인권영향평가가 어떠한 관계가 있는지 모호하다는 것이다.

"인권영향평가가 다른 기준, 법·제도를 준수하고 있는지에 대해 묻는 것보다 다른 가이드에서 명시하고 있는 기준 중에서 인권영향평가에서 꼭 필요한 부분을 준용하여 점검하는 것이 적절할 것으로 보여짐. 현 상황에서는 중복 여부를 따지지 않고, 수많은 가이드라인을 모두 준수하도록 하고 있는 것으로 보여짐" (기업 분야 응답자)

초기 영향평가 과정에서 침해될 우려가 있는 것으로 파악된 인권이 무엇인지에 대한 질의에 대해서는 분야와 상관없이 다수의 응답자가 인공지능이 인권을 침해할 수 있는 구체적인 사례를 제시할 필요가 있다는 의견을 제시했다. 인공지능이 어떻게 인권을 침해할 수 있는지에 대해 평가자가 이해하기 쉽지 않을 수 있다는 것이다. 이와 함께 인권과 '국제인권기준'에 대한 보다 상세한 설명이 필요하다는 의견도 있었다.

인공지능 시스템과 관련된 이해관계자가 누구인지 파악했는지에 대한 질의에 대해, 한 기업 분야 응답자는 "이해관계자 파악은 기업의 영업비밀 유출"이라며 삭제해야 한다는 반응을 보이기도 하였다. 이는 인권영향평가를 불필요한 정부 규제로 보고 경계하는 인식에서 비롯된 오해인데, 이해관계자에 대한 질의가 포함된 취지는 인공지능의 위험을 평가하기 위해 가능한 다양한 이해관계자의 의견을 듣기 위한 것이다. 만일 영업비밀 유출을 우려하여 개발자를 이해관계자에서 배제한다면, 영향평가 과정에 정작 개발자의 관점은 고려되지 않는 결과를 초래할 수 있다.

한편, 한 기술전문가는 '개발자'를 좀 더 엄밀하게 정의해줄 것을 주문했다.

"인공지능 시스템 혹은 서비스와 관련되어 "개발자"는 굉장히 다종다양한 역할, 층위에 있을 수 있습니다.
(1) 인공지능 모델 아키텍쳐의 설계자 및 구현자
(2) 인공지능 모델 훈련(학습)을 위한 데이터 전처리 시스템 개발자
(3) 인공지능 모델 훈련(학습)을 다양한 하이퍼파라메터로 시행하는 개발자
(4) 훈련(학습) 완료된 인공지능 모델이 추론, 예측 혹은 생성(Inference, Prediction, or

Generation)을 실행할 수 있도록 Deploy 하는 개발자
(5) Deploy 된 모델에 대한 추론, 예측, 혹은 생성과 사용자 간의 인터페이스(API 포함) 개발자
(6) 인터페이스에 기반하여 최종사용자가 사용할 수 있는 서비스 혹은 솔루션을 만드는 개발자
(중략) 어떤 개발자를 포함시키고, 어떤 개발자를 포함시키지 않아야 할까요? 혹은 모든 개발자가 포함되어야 한다면 각각의 층위에 있는 개발자가 포함되어야 하는 당위는 어디에 있을까요?" (기술전문가)

3) '조사' 관련 검토항목에 대한 의견

인공지능 인권영향평가를 위한 자료 조사와 관련하여, 한 기업 분야 응답자는 데이터셋과 알고리즘은 모두 기업의 핵심 영업비밀이며 조사 대상과 무관하다고 답변했다. 데이터셋과 알고리즘이 기업의 영업비밀일 수는 있으나 인공지능 시스템의 영향을 평가하는데 필요한 한도에서 이러한 정보에 접근할 수 없다면, 인권영향평가 자체가 불가능해질 수 있다. 따라서 평가자에게 기밀서약을 받는 등 영업비밀을 보호하면서도 인권영향평가를 수행할 수 있는 적절한 방법을 모색할 필요가 있다. 또 다른 기업 분야 응답자는 데이터셋의 목적 및 용도, 장단점에 대한 정보가 문서 형태로 정리되기 쉽지 않다는 한계를 지적하였다.

한 기술 분야 응답자는 '데이터셋'과 '알고리즘'뿐만 아니라, 사전학습모델인 '모델가중치'에 대한 정보가 중요하다며 설명에 이를 고려할 필요가 있음을 강조하였다.

"'데이터셋'과 '알고리즘'보다 '사전학습 모델'(모델 가중치)에 대한 정보가 훨씬 중요할 수 있습니다. 단순히 BERT 알고리즘(모델 아키텍쳐) 사용, 위키백과 데이터셋 사용, 이런 것이 중요한 것이 아니라 이를 기반으로 하여 어느 회사가 어떤 경로로 공개한 사전 학습 모델 사용이 훨씬 중요할 수 있다는 말입니다. 따라서 '데이터셋', '알고리즘' 외에 '모델 가중치'(Model Weight 내지는 Pre-trained Model Weight)에 대

한 정보 또한 매우 중요하다는 것을 드러낼 수 있도록 Q1-3-1의 서술을 보완해주실 수 있다면 매우 좋을 것 같습니다." (기술전문가)

또 다른 기술전문가는 데이터 수집 과정에서 해당 산업군 이해관계자들과의 소통이 필요하다는 점을 제안하였다.

"특정 산업군을 대상으로 하는 데이터 수집 및 인공지능 시스템 개발을 할 경우, 해당 산업군의 대표적인 이해관계자들과 최소한의 설명회 등을 거치는 과정이 필요하지 않을지 생각해봤습니다. 해당 산업군에 대한 이해도가 없는 상태에서 데이터를 수집하고 인공지능 시스템을 개발하여 영향력을 미치려 할 때, 그 생태계 전반이 대상화되는 것은 아닐까 생각이 들기도 합니다." (기술전문가)

나. '2단계' 영향 분석 및 평가에 대한 의견

인공지능 인권영향평가의 2단계 검토항목은 크게 2가지 영역으로 구분된다. 첫째, 인공지능 기술과 관련된 영향분석 및 평가 둘째, 인권에 미치는 영향 및 심각도에 관한 질의이다. 이에 대한 응답자의 주요 의견은 다음과 같다.

1) 인공지능 기술과 관련된 영향 분석 및 평가

'데이터'와 관련하여, 한 기술전문가는 급속하게 변화하고 있는 인공지능 기술 환경 속에서 자칫 질의가 무의미해질 수 있음에 대해 우려를 표명했다.

"학습데이터의 수집과정 및 학습과정 자체가 일견 투명해보이지만 실제로는 사람의 힘으로 도저히 검증할 수 없는 초거대 규모에 달하는 경우, Q2-1-2부터 Q2-1-5까지의 질문에 대해 일괄적으로 '정보 없음'이라고 답하게 될 수 있습니다. 즉, 학습데이터에 대한 검증을 별도로 수행할 수 없는 경우-학습데이터의 전체 공개 없이 사전학습 모델 가중치만 공개된 모델을 사용하여 인공지능 시스템을 구축한 경우- '데이터'와 관련된

영향평가가 조금이라도 유의미성을 지니게 하기 위한 다른 장치가 마련될 필요가 있습니다." (기술전문가)

'알고리즘의 성능과 신뢰성' 과 관련하여, 한 기술 분야 응답자는 '인공지능 시스템은 일정한 위양성, 위음성 비율을 갖고 있을 수밖에 없는데' 라는 표현은 인공지능 시스템의 일부(이진 분류 분야)에만 적용가능하고 회귀나 생성 문제를 다루는 인공지능에는 적합하지 않으므로 수정이 필요하다는 의견을 제시하였다.

한 공공분야 응답자는 "알고리즘의 성능과 신뢰성이 부족하다고 해서 곧바로 인공지능 시스템의 인권침해 요소가 있다고 보기는 어렵지 않을까" 라고 문제를 제기하였으며, 기업 분야 응답자도 "수많은 알고리즘이 있고, 개발자들은 알고리즘을 테스트하고 최선이라고 판단되는 것을 선택하며, 새로운 알고리즘이 지속적으로 개발되고 있는 상황" 이므로 "의도한대로 작동하지 않거나 효율성이 떨어지는 것과 인권과의 관련성은 적은 것으로 보인다" 는 의견을 제시하였다.

'차별금지' 와 관련하여, 각 분야의 응답자가 다양한 의견을 주었다. 질의에 '합리적인 이유없는 차별' 이라는 점을 부연하자는 의견, '경제적 지위, 성별정체성, 병력, 언어' 등 주로 문제 될 사유는 명시하면 좋겠다는 의견 등이 제시되었다.

"예를 들어, 성별 등을 이유로 "특정 집단에 대한 차별을 야기하거나 혹은 기존의 차별을 악화할 가능성" 이 있으면 안 되겠지만, 경우에 따라서 일견 차별에는 해당하나 결국에는 차별로 보지 않는 '예외' 에 해당하는 경우도 있을 수 있습니다. 차별금지 관련 법령이나 지침들도 대부분 원칙-예외 구조로 되어 있기도 합니다. 차별금지 항목뿐만 아니라, 인공지능 인권영향평가도구를 사용하는 주체들은 해서는 안 될 것이 많지만 불가피한 경우도 있지 않냐고 항변할 가능성이 높습니다. 이런 경우에 세심하게 문제를 풀기 위한 길잡이를 어느 정도는 제공해야하지 않을까 합니다." (학술연구자)

한 기술전문가는 구조적 차별을 방지하기 위한 방안으로 단지 개발자의 다양성만이 아니라, 기획자, 디자이너, 마케터, 의사결정자 등 조직 내 모든 구성원을 대상으로 다양성을 고민할 필요가 있다고 제안하였다. 또한, 한 학술연구자는 차별금지와 관련한 질의

를 좀 더 구체화할 필요성을 제기하였다.

“ ‘차별금지’는 어쩌면 인공지능 인권영향평가에서 가장 핵심적인 항목이라고도 할 수 있는데, 질문 하나로 끝나고 있는 점이 아쉽습니다. 위에 열거된 것과 같은 다양한 조건들에 따라 차별이 발생할 수 있는 사례들을 조금 더 구체적으로 언급하면서 그 각각에 대해 검토하였는지 물어보는 문항을 몇 개 만들면 어떨까요” (학술연구자)

‘설명가능성과 투명성’ 관련하여, 여러 응답자들이 추적가능성 및 설명가능성을 현실적으로 담보하기 힘들다는 응답을 하였다. 기업 및 기술 분야 응답자는 “인공지능 시스템을 추적할 수 있어야 한다”는 질의에 대해 아직 현실적으로 추적하는 것이 불가능하다는 의견, ‘인공지능 시스템의 작동 과정을 추적하는 것이 현재 기술 수준에서 불가능한 상황’에서 인권영향평가를 하려고 할 때 기계적으로 ‘아니오’라는 답을 내릴 수밖에 없는 상황을 추동하게 되는 부작용을 낳을 수 있다는 의견을 제시했다. 그러나 이러한 반응 역시 인권영향평가의 모든 질의 항목을 의무적으로 수행해야하는 규제로 보는 인식에 기반하고 있다. 모든 인공지능에 대해 추적가능성을 요구할 필요가 있는 것은 아니라는 점, 반대로 어떤 분야에서는 ‘감사(audit) 추적’을 위해 일정한 추적가능성이 필요하다는 점, 추적가능성은 인공지능 기술에 따라 다르다는 점을 고려할 때 맥락과 상관없이 추적가능성을 요구하려는 것이 아니라, 인권영향평가 과정을 통해 추적가능성이 어느 정도 필요하고 그에 상응하는 조치가 마련되어 있는지 등에 대한 검토를 요구하는 것이라는 취지를 잘 설명할 필요가 있다.

소스코드 공개에 대해서는 오픈소스를 사용하지 않는 경우 검증을 위한 일부 전문가가 아닌 위원회가 필수로 구성되어야 한다는 의견, 소스코드가 공개되었을 때 관련 전문가가 검증할 수 있다고 함부로 말해서는 안된다는 의견 등이 제시되었다.

‘자동화 정도와 인간의 개입’ 관련하여, 한 기업 분야 응답자는 인공지능에 의한 자동화 시스템의 적용을 거부하는 경우, 전적으로 인간의 지원 또는 인공지능이 개입되지 않은 시스템을 대안으로 제공하는지 검토하는 확인 문항이 추가되어야 한다는 의견을 제시하였다. 다른 기업 분야 응답자는 다양한 인공지능 기술들이 서비스에 내재화된 상황에서 모든 인공지능을 이용자에게 알리는 것이 적절한 것인지에 대해서도 고민을 할 필

요가 있다고 제안하였다. 그 외, 법률 분야 응답자는 인공지능 시스템 사용자뿐만 아니라, 운영자도 언제든지 시스템을 정지시킬 수 있도록 해야 한다는 의견을 제시하였다.

‘접근성’과 관련하여, 여러 분야의 응답자들이 경우에 따라 특별한 필요나 장애가 있는 사람들에 대한 인터페이스가 달리 설계·구성될 수 있다는 점을 고려할 필요가 있다는 의견을 제시하였고, 인권단체 분야 응답자는 ‘특별한 필요나 장애가 사람들’이라는 표현을 구체화하여 ‘언어, 나이, 장애, 신체적 조건 등에 상관없이 누구나 설명할 수 있도록 보편적 설계라고 표현” 하는 것이 좋겠다는 의견을 제시하였다.

‘알고리즘 라이선스’와 관련하여, 한 기술 분야 응답자는 알고리즘과 소스코드 등의 용어에 대한 정확한 사용과 서술의 일관성이 필요하다고 지적하며 이 항목은 ‘알고리즘 라이선스’보다 ‘라이선스’로 일반화할 필요가 있다는 의견을 제시해주었다. 한 기업 분야 응답자는 라이선스 판매로 인한 변경범위 및 제한에 대한 명확한 합의가 선행되고, 오픈소스를 활용하는 것에 대해서도 논의가 진행되어야 한다는 의견을 제시하였다. 한 법률 분야 응답자는 변경 권한뿐 아니라 변경 가능한 도구 제공 등 판매자나 사용자가 실질적으로 통제권한을 가질 수 있도록 하는 방안도 검토할 필요가 있다고 제안하였다.

2) 인권에 미치는 영향의 심각도

‘영향을 받는 인권’과 관련하여, 분야 상관없이 다수의 응답자가 질문이 너무 포괄적이어서 좀 더 구체적인 예시가 필요하다는 의견, 그리고 ‘모두’ 파악하는 것은 불가능하다는 의견을 제시하였다. 또한, 앞서 1단계에서 질의한 Q1-2-3과 중복되기 때문에 조정이 필요하다는 의견이 많았다.

한 인권단체 활동가는 한국정부가 비준하지 않은 국제조약은 인권영향평가의 기준이 될 수 있는지 여부에 대해서 명확히 할 필요성을 제기하였다.

“국제 및 국내 인권기준 가운데, 대한민국이 비준하지 않은 국제법은 어떻게 적용하나요? 가령, ‘모든 이주노동자와 그 가족의 권리보호에 관한 국제협약’(약칭 이주노동자권리협약)은 한국이 비준하지 않은 대표적 국제인권법입니다. 평가 자료에 언급한 유엔 시민적·정치적 권리규약(B규약)의 사형제 폐지나 아동권리협약도 비준하지 않았죠.

국제법이지만 정부가 비준하지 않았다는 이유로 평가 기준으로 삼지 않겠다고 할 경우 어떻게 할 것인지요? 이주노동자권리협약에 비춰보면 한국의 고용허가제는 직업선택의 자유가 없는 강제노동의 성격을 갖고 있습니다. 헌법재판소는 이런 제도를 두 번씩이나 합헌 결정을 내렸고요. 인권영향평가는 유엔에 채택한 국제인권규범, 협약 등이 우선해야 한다고 봅니다." (인권단체 활동가)

인권에 미치는 영향의 심각도와 관련하여, 여러 응답자들이 예/아니오의 답변 형식이 아니라, 주관식으로 서술하는 것이 적절하다는 의견을 주었다. 또한, '부정적 영향'의 범위가 모호하다거나 심각도 검토의 객관성을 담보하는 방안이 필요하다는 의견도 있었다. 심각도의 기준을 객관적으로 설정하는 것은 어렵겠지만, 대/중/소 등 대략의 기준을 제시할 필요는 있어 보인다. 또한, 이 질의와 관련해서도 인권영향평가는 객관적인 정답을 산출하는 과정이 아니라, 이해관계자 사이의 협력과 대화를 통해 위험성에 대한 공통의 인식을 형성해나가는 과정이라는 점을 강조할 필요가 있다.

다. '3단계' 방지, 완화, 구제, 이해관계자 협의에 대한 의견

2단계(영향 분석 및 평가)에서 부정적 영향이 파악될 경우, 위험성을 방지하거나 완화하는 조치를 취하고 있는지에 대한 질의를 하였다. 응답자의 주요 의견은 다음과 같다.

'방지' 및 '완화'와 관련하여, 법률 분야 응답자는 잔존하는 위험성에 대해 사용자뿐만 아니라 다른 이해관계자에게도 충분한 정보를 제공할 필요가 있다고 제안하였다. 인권침해 위험을 조속히 탐지할 수 있는 절차나 감독 메커니즘이 제4단계의 인공지능 시스템에 대한 모니터링 여부에 대한 질의와 중복된다는 의견 등 검토항목의 일부 질의가 다른 단계에서의 질의와 중복된다는 의견들이 제시되었다.

'구제'와 관련하여, 여러 응답자들이 권리구제의 절차 마련과 정보 및 신청 절차가 명확하게 공개되는지에 대한 의견을 주었다. 기업 및 법률 분야 응답자는 "결정에 대한 이의나 권리구제 요구를 적극적으로 수용할 수 있는 인센티브 체계가 마련되어 있는지, 또는 이와 같은 요구의 불수용을 유도할 유인이 적절히 식별되고 통제되고 있는지 항목으로 보충이 필요"하다는 의견, 그리고 권리구제 절차 마련과 더불어 '이의를 제기할

수 있는 기관과 방법에 대한 정보가 일반에 공개되어 있는지’도 함께 확인할 필요가 있다는 의견이 있었다.

라. ‘4단계’ 공개 및 점검에 대한 의견

4단계 공개 및 점검에서는 인공지능시스템의 주요 요소 공개, 인권영향평가 결과 공개, 인공지능시스템에 대한 모니터링, 인권영향평가에 대한 점검, 인권영향평가의 반복적 수행에 대한 질의를 하였다. 응답자의 주요 의견은 다음과 같다.

인공지능 시스템의 주요 요소의 공개 관련하여, 우선 앞서 2단계에서 검토했던 설명가능성 및 투명성 질의와 중복된다는 의견이 제시되었다. 다만, 2단계에서는 이해관계자에의 설명에 방점을 두었다면 4단계에서는 일반 공개에 초점을 두었다는 차이가 있다.

한 기업 부문 응답자는 공개의 범위가 모호하다는 의견을 표명하였다.

“많은 기업들이 인공지능 시스템에 대해서 이용자에게 알리기 위한 노력을 하고 있음. 이러한 상황에서 데이터와 알고리즘의 주요 요소를 일반에 공개하고, ‘이해할 수 있는 방식으로 쉽게 설명’하는 것의 범위가 매우 추상적이어서 엄격하게 적용할 경우 이를 준수할 수 있는 사업자는 거의 없을 것으로 보여짐” (기업부문 응답자)

이와 관련하여 국제적으로 알고리즘의 주요 요소의 공개를 요구하는 규정이 도입되고 있어 점차 구체화된 관행이 수립될 가능성이 크다는 점, 그리고 주요 요소 공개의 취지를 공감한다면 가능한 수준에서 해당 조직이 공개할 수 있는 주요 요소를 검토하는 계기를 부여하는 것이 인권영향평가의 기본 취지라는 점을 환기할 필요가 있다. 이는 법률을 통해 알고리즘의 주요 요소의 공개를 의무화하는 것과는 다르다.

‘인권영향평가 결과 공개’와 관련하여, 일부 학술 및 기업 분야 응답자는 공공영역과 달리 민간영역에서는 인권영향평가 보고서를 공개하는데 충분한 검토와 준비가 필요하다는 의견, 모든 사항을 문서화하여 보고서에 담는 것은 사업자가 이행하기 어려우니 어느 정도까지 공개하는 것이 적절한지에 대한 추가 검토가 필요하다는 의견을 제시하였다. 다만, 공공영역에서 인권영향평가 결과를 공개하는 것에 대해서는 별다른 이견이 없

었다.

“공공영역에서 사용하는 인공지능은 전 국민에게 끼치는 영향이 크기 때문에 반드시 결과를 공개하여야 한다.” (인권단체 활동가)

‘인공지능시스템에 대한 모니터링’ 관련해서는 전반적으로 큰 이견은 없었다.

“인공지능시스템이라는 전문영역을 모니터링하는 것은 이해관계자도 어렵습니다. 이는 개발자나 운영자의 협조와 국가인권위의 적극적인 개입 의지가 필요한 부분입니다. 이해관계자에 대한 모니터링 교육도 있었으면 합니다” (인권단체 활동가)

‘인권영향평가에 대한 점검’과 관련하여, 기업 및 공공분야 응답자는 인권영향평가 수행의 효과와 한계를 점검하는 절차에 더하여, 이의 개선을 위한 절차까지 마련하도록 할 필요가 있다는 의견을 제시하였다.

‘인권영향평가의 반복적 수행’과 관련하여, ‘반복적’이라는 표현보다는 ‘재수행’이 바람직하다는 의견, 주기적 수행은 많은 노력이 필요하므로 가급적 특정한 경우 재실시하도록 하는 것이 바람직하다는 의견 등이 제시되었다.

3. 인공지능 인권영향평가(안)에 추가적으로 포함해야 할 질의 분야 및 항목에 대한 의견

인공지능 인권영향평가도구에서 추가적으로 포함해야 할 질의 분야 및 항목이 있다면 제안해주시기 바랍니다.

한 법률 분야 응답자는 “공공기관에서 사용하는 인공지능의 경우 그 영향이 크기에 직접적인 이해관계인 외에도 시민사회에서 관련 계획 및 개발 과정에 참여하고 의견을 내는 공청회 등의 절차가 마련될 필요가 있다” 고 제안하였다.

어느 학술 분야 응답자는 인공지능 인권영향평가가 수행된 이후에 그것이 “다시 조직 내에 환류되어 영향평가라는 특별한 절차를 밟지 않고도 인공지능 개발의 전 과정에서 일상적으로 수행될 수 있도록 하는 방안”에 대한 고민이 필요하다는 것, 또한 “4단계 마지막에 있는 ‘반복적 수행’을 넘어서 조직 내에 이 문제가 공유될 수 있도록 하는 방안을 제시하면 좋겠다”는 의견을 제시해주었다. 인권영향평가의 제도화 과정에서 지속적으로 평가가 필요한 부분이다.

4. 인공지능 인권영향평가(안) 전반에 대한 추가의견

> 귀한 시간 내주셔서 감사합니다. 마지막 질문으로, 인공지능 인권영향평가(안) 전반에 대한 추가 의견을 제시해주시기 바랍니다.

한 학술 분야 응답자는 영향평가의 대상을 구체화하고, 자신이 개발 운영하는 인공지능이 대상이 되는지 판단하기 위한 ‘체크리스트’를 만들어 제시하면 좋겠다는 의견을 제시했다. 본 인권영향평가는 기본적으로 독립적인 평가자에 의한 평가를 전제하지만, 개발자 역시 본 도구안을 자체 평가 목적으로 활용할 수 있다.

어느 기업 분야 응답자는 전체적으로 다양한 주제를 다루고 있어 각 단계와 항목별 ‘연계’가 분명하게 드러나도록 구성하여 ‘체크리스트’ 작성 과정에서 연계가 있는 단계와 항목을 종합적으로 살펴보면 좋겠다는 의견을 제시하였다. 또한 평가도구 도입부에 간단한 인포그래픽을 제시하는 방안도 제안하였다.

몇몇 응답자들은 인공지능 인권영향평가가 기업들에게 규제로 느껴질 수 있다는 점을 지적하였다. 한 학술 분야 응답자는 인공지능 인권영향평가의 효과적인 수행을 위해서는 공감할 수 있는 친절한 설명이 필요하다고 의견을 제시하였다. 인공지능 시스템 개발자나 관리자가 인권문제에 과도하게 추궁당한다는 느낌을 받으면 인권영향평가가 효과적으로 실행되기 어렵다는 의견이다. 관련해 기업 분야 응답자는 해당 평가에 대한 취지에는 공감하나 기업의 경우 영향평가나 지침이 규제로 받아들여진다는 것, 문제가 발생할 경우 이를 개선하여 더 좋은 서비스를 만드는 것에 집중하는 것보다 책임이 누구에게 있

는지를 검토하는 것으로 비춰질 수 있다는 점에서 신중한 논의가 필요하다는 의견을 제시하였다. 인권영향평가가 형식적인 규제가 아니라 인공지능의 위험성을 이해관계자와의 소통에 기반하여 사전에 평가함으로써 오히려 사회적인 갈등과 비용을 줄이는 방법이 될 수 있도록 인권영향평가의 의미에 대해 더 많은 소통과 논의가 필요한 것으로 보인다.

한편, 공공분야 응답자는 영향평가의 특성상, 그리고 인공지능의 특성상, 평가가 필요한 부분이 광범위하기 때문에 문항이 많은 것은 당연하지만, 점검 문항의 우선 순위를 두고 문항 수를 줄이거나 중복문항을 통합 및 삭제하는 방법을 고려하여 평가의 실효성을 높여야 한다는 의견을 제시하였다.

인권단체 분야 응답자 다수는 인공지능 시스템의 개발과 활용이 확대되므로 인공지능 인권영향평가도구가 인권침해의 사각지대 영역을 파악하는데 도움이 될 수 있어야 한다는 의견, 구제 절차의 명확한 지침을 구축할 수 있는 계기가 되길 바란다는 의견 등을 제시하였다. 또한, 인공지능 활용정책과 사업에 따라 좀 더 집중적으로 검토, 확인할 사안이 존재할 것인데, 특히 돌봄영역에서 인공지능 시스템 개발과 활용과 같은 영역에서 이를 보완할 수 있는 영향평가도구가 마련되어야 한다는 의견이 있었다. 앞으로 인공지능 인권영향평가가 제도화되고 폭넓게 활용된다면, 인공지능 적용 분야별로 인권영향평가도구가 세분화될 필요도 있을 것이다.

학술 분야 응답자는 두 가지 의견을 제시하였다. 정량적인 위험도 측정 등을 문서화하는 것이 필요하다는 것과 인공지능 인권영향평가 법령을 통해 의무화되어야 의미가 있다는 의견이 있었다.

“고위험 인공지능 중 “필요한 경우에는” 정량적인 위험도 측정 등을 문서화하도록 하면 좋지 않을까 싶습니다. 예를 들며, 가명처리의 익명화 수준 등을 측정하여 이를 보고서 등 세부 항목에 기재토록 하는 방안의 검토가 필요할 것으로 보입니다. 그 이유는 통상 인공지능 알고리즘에 수반된 위험이 정도(degree)의 속성을 가지기 때문입니다. 또한, 인공지능 인권영향평가의 경우, 다른 국내외 제반 영향평가들과 마찬가지로 법령을 통해 의무화되어야 의미가 있을 것으로 판단됩니다. 따라서 관련 연구를 추진하심에 있어 단순 권고 수준이 아니라 법령상 의무사항으로 할 수 있는 제도적 방안도 제시해주시면 좋을 것 같습니다.” (학술연구자)

제5장 인공지능 인권영향평가 도입 방안

제1절 제도적 형식의 측면

인공지능에 대한 인권영향평가를 시행하게 되면 인공지능기술이 개발 또는 도입되기 전, 나아가 인공지능기술을 활용한 사업 및 정책이 시행되기 전에 인권적 관점에서 그 영향을 파악할 수 있고, 이로 인한 부정적 영향을 완화하거나 제거할 수 있으며, 긍정적인 영향은 극대화해나갈 수 있다. 또한 인권에 미치는 영향을 파악함으로써 그 위험을 관리해나갈 수도 있다.

그러나 현재 인공지능기술 및 인공지능기술을 활용한 사업 또는 정책을 대상으로 한 인권영향평가가 실질적으로 제도화되었다고 보기는 어렵다. 2020. 12. 10. 시행된 「지능정보화 기본법」 제56조에서 국민의 생활에 파급력이 큰 지능정보서비스 등의 사회적 영향평가에 대해 정하면서, 지능정보 서비스 활용에 대해 안전성 및 신뢰성, 이용자의 권익이나 정보보호에 미치는 영향에 대해 평가하도록 정하였으나, 인권에 미치는 부정적 영향을 식별·평가하는 인권영향평가와는 목적과 방향성을 달리한다. 또한 지능정보서비스의 사회적영향평가에 관한 규율은 실질적으로 영향평가를 추진하기 위한 영향평가의 체계, 범위 및 절차에 관한 구체적인 근거가 마련되어 있지 않고, 추상적, 선언적인 수준에 머물러 있다.

과학기술기본법 제14조 및 동법 시행령 제23조에서 미래의 신기술 및 기술적·경제적·사회적 영향과 파급효과 등이 큰 기술에 대하여, 해당 기술이 국민생활의 편익증진 및 관련 산업의 발전에 미치는 영향, 새로운 과학기술이 경제·사회·문화·윤리 및 환경에 미치는 영향, 해당 기술이 부작용을 초래할 가능성이 있는 경우 이를 방지할 수 있는 방안, 해당 기술의 성격과 파급효과가 성별 등 특성에 미치는 영향에 대하여 기술영향평가를 실시하도록 정하고 있으나, 이 역시 인권보호의 측면에서 이뤄지는 영향평가제도로 보기는 어렵다.

인공지능기술의 영향력이 급속도로 확대되어 가는 상황에서 인권에 미치는 영향을 파악하고, 그로 인한 피해를 방지하기 위해 인공지능기술에 대한 인권영향평가제도의 실질

적인 제도화가 시급하다고 할 것이다.

이하에서는 인공지능 인권영향평가의 법적 근거 마련 방안에 대하여 살펴보고, 나아가 제도화의 주요 형식에 대하여 검토해본다. 인공지능 인권영향평가도 인권영향평가에 속하므로 인권영향평가와 관련한 기존의 논의와 경험을 주되게 참고하면서 인공지능 기술의 특성을 고려하여 논의하여 보고자 한다.

1. 인공지능 인권영향평가의 법적 근거 마련[120]

인공지능기술을 개발, 도입하거나 인공지능기술을 활용한 사업 또는 정책을 시행함에 있어 인권영향을 식별 및 평가하기 위해 거쳐야 할 절차를 명확하게 설계하고, 일정한 의무를 부여하며, 공개 및 검증을 거치도록 하기 위해서는 인공지능기술에 대한 인권영향평가 시행의 법률적 근거를 명확하게 마련할 필요가 있다.

일부 지방자치단체에서 개별적인 조례에 근거하여 일반적인 인권영향평가를 시행하는 사례가 있으나, 실효성 있게 실시되고 있다고 평가하기 어렵고, 각 지방자치단체별로 개별적인 기준에 따라 시행되고 있어 일률적인 기준이 부재한 상황이다. 인공지능기술과 관련한 표준화된 인권영향평가를 시행하기 위해서는 다른 사회영향평가와 마찬가지로 정부차원에서의 지침이 마련되어야 하고, 이를 위해서는 법률적 근거가 마련되어야 할 것이다.

그 구체적인 방안으로 먼저 기존의 사회영향평가에 인공지능 인권영향평가를 포함시키는 방안을 고려해볼 수 있다. 우리나라의 영향평가제도는 통합적 영향평가 방식이 아닌 다양한 영역에 대한 개별적인 영향평가 방식을 취하고 있다.[121] 이에 개인정보영향평가, 환경영향평가, 성별영향평가, 기술영향평가 등 영향평가의 개별 근거법령상 세부 평가 항목에 인공지능기술이 인권에 미치는 영향에 관한 항목을 추가시켜 기존의 법률을 개정하는 형태로 인공지능기술에 대한 인권영향평가제도의 법률적 근거를 마련하는 방안을 고려해볼 수 있을 것이다.

이는 불필요한 비용의 지출을 줄이고, 효율성을 도모할 수 있다는 장점을 지닌다. 이

120) 이 부분 논의의 전체적인 구조는 인권영향평가제도의 법적근거 마련과 관련한 이충은 외(2018)의 내용을 참고하였다.
121) 이준일 외(2018), 98면 참조.

미 시행되고 있는 기존의 영향평가의 위험 관리구조를 활용할 수 있고, 여러 방면의 평가를 반복적으로 했을 때의 피로를 방지할 수 있으며, 개인정보보호, 사생활보호, 환경보호, 부패방지 등과 같은 인권영향의 상호 관련성에 대한 분석을 촉진할 수 있다.

그러나 기존의 개별 사회영향평가제도는 각 제도별로 소관부처를 달리하고, 영향평가를 통해 추구하는 목적에도 차이가 있다. 개인정보보호위원회가 소관하는 개인정보영향평가제도는 개인정보 침해의 위험요인 분석과 개선사항 도출을 목적으로 하고 있어 인권과 관련성이 있지만 개인정보 침해와 관련한 영향분석이 주를 이룬다는 점에서 인권의 일부분에 해당될 뿐이다. 환경부 소관인 환경영향평가제도는 그 평가항목에 사회·경제적 영향평가 항목을 포함하고 있지만 인권영향평가와는 거리가 멀다.[122]

개별 사회영향평가제도의 평가항목에 인권영향평가에 관한 항목을 포함시키는 것만으로는 고유의 목적을 실현하기에 어려움이 있고, 인권침해의 가능성이나 위험을 소홀히 다룰 우려가 존재한다. 따라서 기존의 사회영향평가에 인공지능 인권영향평가제도를 결합시키는 방안은 바람직하다고 보기 어렵다.

두 번째로 국가인권위원회법 개정을 통해 인공지능 기술 관련 인권영향평가의 법률적 근거를 마련하는 방안을 고려해볼 수 있다.

과거 2003. 12. 1.경 16대 국회에서 천정배의원 대표 발의로 국가 또는 지방자치단체가 인권의 보호와 향상에 영향을 미치는 법령·정책 등을 제정하거나 입안하고자 할 때 인권영향평가서를 작성하여 국가인권위원회에 제출하도록 의무화하고, 평가서를 작성할 때 설명회 또는 공청회를 개최하여 의견을 수렴하도록 하며, 문제가 있을 경우 국가인권위원회가 해당 기관에 법령, 정책 등의 중단을 권고할 수 있도록 하는 내용의 법안(의안번호 : 2978)이 발의되었으나, 동 법안은 2004년 5월 국회 임기만료에 따라 자동폐기되었다.

이와 마찬가지 방식으로 국가인권위원회법 개정을 통해 인공지능 인권영향평가 제도 도입의 법률적 근거를 마련하고, 시행령, 시행규칙의 형태로 영향평가의 체계, 절차, 방식을 구체화시킬 수 있을 것이다. 다만 이 방안은 다시 상정되어 통과되기까지 적지 않은 논의와 노력이 필요할 것으로 예상된다.

세 번째로 새로운 법률을 제정하는 방안이 있다. 법무부와 국가인권위원회는 인권정책

122) 이충은 외(2018).

의 통합적이고 효율적인 추진을 위하여 공동으로 인권기본법안의 입법을 추진해왔고, 2021. 12. 28.경 인권정책기본법 제정안이 국무회의를 통과하였는데, 위 제정안 제23조에 의하면 정부는 기업의 인권존중책임에 대하여 평가기준 및 평가지표를 설정·운영할 수 있고, 이와 관련하여 국가인권위원회에 의견 및 협력을 요청할 수 있도록 하였다. 아주 명확한 형태는 아닐지라도 이는 인권영향평가 제도 시행의 근거로 볼 수 있다.

또한 인공지능 기술의 광범위한 활용성과 영향력을 고려할 때, 인공지능 기술에 대하여 독자적인 형태로 인권영향평가법을 제정하는 방안도 고려해볼 수 있을 것이다. 다만 위와 같은 새로운 법률의 제정에는 마찬가지로 적지 않은 논의와 노력을 필요로 할 것으로 예상된다.

정리하면, 인공지능 인권영향평가는 기존에 시행되고 있는 개별 영향평가제도에 결합하는 형태보다는 인권영향평가만의 고유의 목적을 실현할 수 있도록 단독으로 실시되는 형태가 바람직하다고 할 수 있다. 따라서 다른 영향평가의 근거 법령을 개정하는 형태보다는 국가인권위원회법을 개정하거나 인권기본법이나 인공지능기술에 대한 독자적인 인권영향평가법 등 새롭게 법안을 제정하는 형태가 적절하다고 할 수 있다. 다만, 법개정이나 법률 제정 등의 방식은 많은 논의와 노력, 그리고 시간을 필요로 하기 때문에 현 시점에서 이를 전제로 논의하기에는 다소 시기상조라 할 수 있다.

비록 명시적인 법률적 근거는 부재한 상황일지라도, 현 상태에서도 하위 법령 수준에서 인공지능 인권영향평가제도 시행의 법적 근거 자체는 마련되어 있다고 할 수 있다.

대법원은 2015년경 전라북도 학생인권조례를 대상으로 한 조례안의결무효확인의 소[123]에서 “이 사건 조례안은 전체적으로 헌법과 법률의 테두리 안에서 이미 관련 법령에 의하여 인정되는 학생의 권리를 열거하여 그와 같은 권리가 학생에게 보장되는 것임을 확인하고 학교생활과 학교 교육과정에서 학생의 인권 보호가 실현될 수 있도록 그 내용을 구체화하고 있는 데 불과”하다고 하면서 “이 사건 조례안 규정들이 헌법과 관련 법령에 의하여 인정되는 학생의 권리를 확인하거나 구체화하고 그에 필요한 조치를 권고하고 있는 데 불과한 이상 그 규정들이 교사나 학생의 권리를 새롭게 제한하는 것이라고 볼 수 없으므로, 국민의 기본권이나 주민의 권리의 제한에 있어 요구되는 법률유보원칙에 위배된다고 할 수 없고, 그 내용이 법령의 규정과 모순·저촉되어 법률우위원칙에

123) 대법원 2015. 5. 14. 2013추98 판결 [조례안의결무효확인]

어긋난다고 볼 수도 없다.” 고 판시하였다.

이로 미루어 인권영향평가제도와 같이 헌법에서 인정되는 권리를 구체화하고, 직접적으로 인권을 제약하는 국가행위라고 할 수 없는 제도는 직접적인 법률의 위임없이도 도입할 수 있다는 점을 확인할 수 있다.[124]

그리고 이 지점은 국가인권위원회의 적극적인 역할을 기대해볼 수 있도록 하는 부분이다. 국가인권위원회는 인권에 관한 법령·제도·정책·관행의 조사와 연구 뿐만 아니라 그 개선이 필요한 사항에 관하여 권고 또는 의견을 표명할 수 있고, 인권침해의 유형, 판단기준 및 그 예방조치 등에 관한 지침을 제시 및 권고할 수 있으므로, 인공지능에 대하여 국가인권위원회가 인권의 보호와 향상을 위한 업무로서 얼마든지 인공지능기술 또는 인공지능기술을 활용한 사업이나 정책에 대하여 인권보호의 관점에서 모범적인 지침 내지 권고안을 마련하여 그 시행을 권고하거나 의견을 표명할 수 있다.[125]

이하에서는 인공지능기술의 개발 및 도입, 관련 사업 및 정책의 도입에 따른 인권영향을 식별, 평가하고, 인권 침해를 예방하거나 침해를 구제할 수 있는 실효성 있는 인권영향평가제도가 되기 위한 제도의 주요 형식을 검토해보고자 한다.

2. 제도화의 주요 내용

가. 평가 대상

인공지능 기술이 매우 다양하고 광범위하게 활용되고 있고 향후 그 정도가 더욱 심화할 것임을 고려할 때 모든 인공지능 기술 및 이를 활용한 사업 또는 정책을 인권영향평가의 대상으로 삼을 수는 없다. 이에 인권영향평가의 대상을 일정한 기준에 따라 한정할 필요가 있는데, 이는 인권영향평가 시행 및 그 결과에 대한 구속성 등 의무성 부과와 결부지어 고민해보아야 하는 문제이다.

---

124) 이준일 외(2015), 77면 참조. 한편, 이는 일부 지방자치단체가 상위법의 직접적인 위임 없이도 조례의 형식으로 인권영향평가제도를 도입하여 시행할 수 있는 법적 근거로도 볼 수 있다.

125) 한편 앞서 언급한 지능정보화 기본법 제56조는 인공지능기술 및 그 서비스의 사회적 영향평가에 관한 법률의 근거를 추상적인 수준에서나마 마련하고 있는데, 동 규정에 의하면 국가 및 지방자치단체를 그 시행주체로 정하고 있어, 해당 법조항이 인권 전담 국가기관인 국가인권위원회의 적극적인 역할 수행에 대한 법률적인 근거가 될 가능성도 없지는 않다.

해외의 사례를 보면, 대체로 인공지능 시스템의 위험도를 평가하여 위험도별로 조치를 달리 차등 적용하고 있음을 알 수 있다. 유럽연합의 인공지능법(안)은 인공지능 시스템을 수용불가한 위험, 고위험, 제한된 위험, 최소한의 위험으로 구분하여, 각 시스템별로 의무사항을 달리 정하고 있다.

이는 위험수준에 따라 요구사항을 차등 적용하는 위험기반접근법에 따른 것인데, 앞서 언급한 바와 같이 인권기반접근법에 따르는 경우에도 침해의 심각도를 고려하고 심각성이 높을수록 더욱 신속하고 엄격한 조치를 수반하도록 정하고 있어 구체적인 표현만을 '심각도'로 달리하고 있을 뿐, 인권에 미치는 영향에 따라 요구사항을 차등 적용한다는 점에 있어서 차이가 없다고 할 수 있다.

국가인권위원회의 인공지능 개발과 활용에 관한 인권 가이드라인에서도 인공지능이 금지되는 영역, 상당한 제한이 필요한 인공지능 고위험 영역, 위험성이 거의 없는 영역 등 적절하게 위험성 단계를 구분하고, 그에 맞는 규제 수준과 인적 개입이 이루어지도록 법과 제도를 마련할 것을 권고하였다.

이에 위험도에 따라 의무 부과 및 요구사항을 차등적용하는 방식의 합리성은 이미 어느 정도 확인된 상태라고 할 수 있는데, 이 경우에도 구체적으로 어떤 기준에 따라 위험도를 측정, 평가하여 인권영향평가의 대상으로 삼을 수 있을지의 문제는 여전히 남는다.

그런데 우선 그 자체로 법률에서 금지하는 인권 침해 또는 차별 대우를 목적으로 하거나 법률에서 금지하는 개인정보의 처리를 목적으로 하는 인공지능의 경우 수용할 수 없는 수준의 매우 심각한 인권 침해의 위험이 존재한다고 할 수 있으므로 이는 일응 '금지되는 인공지능'으로 분류하여 그 개발이나 활용을 즉각적으로 중단하도록 의무화할 것을 권고하는 것이 바람직하다고 할 수 있다.

예를 들어, 공공장소에서 대량 감시와 차별로 이어지고, 집회 및 결사의 자유에 부정적 영향을 나타낼 위험이 높은 얼굴인식 등 원격 생체인식을 목적으로 하는 인공지능, 생명의 존엄성 및 윤리를 훼손할 가능성이 높은 자율살상 무기에 활용되는 인공지능은 그 자체로 수용하기 어려운 심각한 수준의 인권 침해 위험성을 지니고 있다고 평가할 수 있다.

유럽연합의 인공지능법(안)은 ① 사람의 의식을 뛰어넘는 잠재의식 기술을 배치, 인지하지 못하는 방식으로 인간의 행동, 의견 또는 결정을 조작하여 자신 또는 타인에게 신

체적·정신적 위험을 가져올 수 있는 인공지능 시스템, ② 개인·단체에 대한 정보 및 예측을 악용하여 아동·장애인 등이 취약성 또는 특수 상황을 표적으로 삼는 인공지능 시스템, ③ 공공기관이 사회적 행동 또는 알려지거나 예측된 개인의 특성을 기반으로 자연인의 신뢰도를 평가하거나 사람의 특성을 분류하여 불리한 대우를 하는 인공지능 시스템, ④ 경찰 등이 공개된 장소에서 실시간으로 생체정보를 활용하여 신원확인을 하는 인공지능 시스템 등의 경우 금지되는 인공지능시스템으로 분류하였는데, 이와 같이 도입 또는 적용되는 분야를 기준으로 금지되는 인공지능을 분류하는 방식은 참고할만하다. 다만 금지되는 인공지능의 구체적인 분야를 현단계에서 확정적으로 제시하기에는 한계가 있고, 이에 대하여는 향후 다양한 사회적 논의와 의견 수렴이 이루어져야할 것으로 보인다.

금지되는 인공지능 외에도 위험도가 높은 인공지능의 경우 이를 일응 '고위험 인공지능'으로 분류하여 보다 강한 의무와 요구사항들을 부과하는 방향이 합리적인데, 이때 위험도의 측정 및 평가에는 양적인 지표와 질적인 지표를 동시에 고려하여야 할 것이다.

인공지능 개발과 활용에 따라 영향을 받는 당사자의 수, 사용된 데이터의 양 등이 많을수록 인권 침해의 가능성이 객관적으로 높아진다고 할 수 있으므로, 이를 반영한 양적 기준에 따라 대상 범위를 한정하되, 인권침해와 차별의 가능성 및 정도와 관련한 질적 기준도 함께 고려하여 인공지능 시스템의 위험도를 객관적으로 평가하는 것이 바람직하다. 기존 제도로 관리되거나 감독될 수 없는 새로운 분야 또는 개인의 생명이나 안전 등 인권에 중대한 영향을 미치는 분야에 인공지능 기술이 개발, 적용되는 경우는 양적 기준과는 무관하게 고위험으로 분류하여 영향평가의 대상으로 삼을 수 있을 것이다.[126]

유럽연합의 인공지능법(안)의 경우 제품안전, 생체인식 및 분류, 중요 인프라의 관리 및 운영, 교육 및 직업훈련, 고용 및 직원 관리, 중요한 민간 및 공공 서비스의 접근과 이용, 법집행, 이민 및 국경통제, 사법 업무 등에 관련된 인공지능 시스템을 고위험 인공지능으로 분류하고 있는데, 이는 도입 또는 적용되는 분야를 기준으로 인권침해의 잠재적 위험성을 구분한 것으로 그 방식과 내용은 참조할만한 유용한 예시 사례라 할 수 있

126) 국가인권위원회(2022) 가이드라인에서도 인권침해와 차별의 가능성 및 정도, 영향을 받는 당사자의 수, 사용된 데이터의 양 등을 고려하여 인권영향평가를 실시해야하고, 기존에 관리되거나 감독될 수 없는 새로운 분야는 인권영향평가를 도입해야한다고 설명하고 있다.

을 것이다.

미국 의회가 2022. 2. 3. 발의한 알고리즘 책무성법(안)은 평가, 검·인정 또는 인증을 포함한 교육 및 직업훈련, 고용, 노동자 관리 또는 자영업, 전기, 난방, 수도 인터넷 등 정보통신망 접근 또는 교통 등 인프라의 운영, 입양 또는 출산 관련 서비스를 포함한 가족계획, 금융서비스, 의료서비스, 주거 관련 서비스, 법률서비스 등 영역과 관련하여 이루어지는 결정, 판단을 '중요한 의사결정'이라고 정의하면서, 위 '중요한 의사결정'에 활용되는 자동화된 의사결정 시스템을 위해 개발되거나, 같은 시스템에 활용될 것이라고 합리적으로 예측가능한 자동화된 의사결정 시스템에 대하여 영향평가를 실시하고, 운영과정에서 관련된 정보의 기록, 보관의무 등을 정하고 있는데, 이 역시 위 유럽연합의 인공지능법(안)과 마찬가지로 도입 또는 적용된 분야를 기준으로 위험성의 정도를 구분한 사례이다[127].

2021. 7. 1.경 정필모 의원의 대표발의로 제안된 <인공지능 육성 및 신뢰 기반 조성 등에 관한 법률안>에서는 의료, 전기·가스·수도·핵시설 등 에너지, 기간서비스, 개인에 대한 평가 또는 의사결정, 공공기관 사용 부문에서 사람의 생명·신체에 위험을 줄 수 있거나 부당한 차별 및 편견의 확산 등 인간의 존엄성을 해칠 위험이 있는 인공지능을 '특수 활용 인공지능'으로 정의하면서(제2조 제2호) 의사결정 원리 및 최종결과 등을 설명하도록 하거나(제20조 제3항) 기술적·관리적 조치를 하도록 하도록 하는 등(제21조) 부가적인 의무를 부과하였는데, 이 역시 유럽연합의 인공지능법(안)과 미국의 알고리즘 책무성법(안)과 마찬가지로 인공지능 기술이 도입 또는 적용된 분야를 기준으로 잠재적인 위험도를 달리 평가하는 구분방식을 따른 것이다.

마찬가지로 2021. 11. 24.경 윤영찬 의원의 대표발의로 제안된 <알고리즘 및 인공지능에 관한 법률안> 제2조 제3호에서는 인간의 생명, 생체인식, 교통·수도·난방·전기 등 주요 사회기반시설, 채용 등 인사평가 또는 직무배치의 결정, 응급서비스, 대출 신용평가 등 필수 공공·민간 서비스, 수사 및 기소 등 권한행사, 출입국 관리 등 분야에 사용하는 인공지능을 '고위험인공지능'으로 정의하면서 위험관리시스템 구축, 이용자에 대한 정보제공 등 부가적인 의무사항을 규정하고 있다.

---

127) 한국지능정보사회진흥원 정책본부 지능화법제도팀(2022). 미국, '알고리즘 책임법안(Algorithmic Accountability Act)' 발의. 디지털 법제 Brief(2022. 3. 11). 한국지능정보사회진흥원(NIA).

이처럼 도입 또는 적용되는 분야를 기준으로 인공지능의 위험도를 구분하는 형태는 다수의 선행 사례 뿐만 아니라 국내에서 발의된 법안에서도 활용되고 있으나, 각 경우에 위험도가 높다고 평가한 구체적인 대상 분야에는 적지 않은 차이가 있다는 점 또한 확인할 수 있다.

이에 현 단계에서 이를 확정적으로 정하기에는 무리가 있고, 향후 다양한 사회적 논의와 의견 수렴을 통해서 고위험 인공지능을 명확히 규정하려는 시도가 이루어져야 할 것으로 본다. 다만 유럽연합의 인공지능법(안)과 미국의 알고리즘 책무성법(안) 및 국내에서 발의되고 있는 법안에서는 군사 목적으로 개발되거나 사용하는 인공지능과 정보기관이 사용하는 인공지능을 인권 침해의 잠재적인 위험성이 높은 분야의 범주에서 제외하고 있는데, 위 각 분야에 활용되는 인공지능의 경우 오히려 높은 인권 침해의 가능성을 내포하고 있다고 할 수 있어, 달리 배제시킬 이유를 찾기 어렵다는 점은 지적되어야 할 것이다.

한편 위의 논의와는 별개로 공공분야에서 인공지능을 개발하거나, 조달하여 활용하는 경우 그 자체로 민간 분야에 비해 파급력이 크다고 할 수 있다. 공공기관은 민간 기업에 비하여 보다 두터운 공정성, 합법성, 책무성의 부담을 지니므로 일응 위험도와 무관하게 공공기관이 개발하고 활용하는 모든 인공지능을 영향평가의 대상으로 삼을 필요가 있어 보인다.

나. 시행 시기

앞서 영향평가제도의 유형에서 살펴본 바와 같이 인권영향평가제도는 수행시점에 따라 ‘사전적’ 영향평가와 ‘사후적’ 영향평가로 구분될 수 있는데, 인권침해의 위험이 큰 인공지능기술이 현실에 실제로 적용된 이후에는 이미 인권에 대한 부정적 영향, 즉 인권 침해가 발생하여 사후에 영향평가제도를 통해 인권침해의 위험을 식별, 평가하더라도 뒤늦은 조치가 될 가능성이 크고, 사후적 구제에는 많은 비용과 시간이 소요된다.

특히 인공지능 알고리즘 개발의 경우 프로젝트의 진행 정도가 성숙해질수록 문제점을 개선하기 어려운 특성이 있다. 이에 프로젝트의 설계 초기 단계에서 인공지능 알고리즘의 개발 자체가 부적절하거나 바람직하지 않은 것으로 판명되는 경우 즉각적으로 문제를

수정하거나, 프로젝트의 진행을 중지하여야만 비용과 시간의 낭비를 방지할 수 있다.

따라서 인공지능기술에 대한 인권영향평가는 인공지능기술이 실제로 개발되어 적용되기 이전에 사전적으로 실시하여 인권에 대한 부정적 영향을 예측, 예방하는 것이 필수적이라고 할 수 있다. 이때 사업의 내용이 구체화되기 이전에는 영향평가를 실시하는 것이 무의미할 수 있으므로 적어도 인공지능기술 개발 또는 도입에 관한 구상이 충분히 구체화된 시점 이후에 실시하도록 하는 것이 바람직하다.

다만 인공지능기술의 특성을 고려할 때 사전에 시행되는 인권영향평가에 한정하지 않고, 정기적, 사후적 평가를 병행하여 지속적으로 그 인권영향을 식별하고 평가할 필요성이 있다. 특히 사전영향평가만으로는 기존에 이미 도입되거나 시행되고 있는 인공지능시스템으로 인하여 이미 발생한 인권 침해나 차별을 시정하거나 구제하는 것이 불가능하다.

또 인공지능시스템의 기반 기술은 계속해서 진화하고 있을 뿐 아니라, 동일한 기술이라 하더라도 인공지능시스템이 운영되는 지정학적, 사회적, 경제적 맥락에 따라 위험성이 달라질 수 있다.[128] 이러한 인공지능의 특성을 고려할 때 사전에 시행되는 영향평가만으로 이로 인한 인권침해의 가능성이나 위험을 충분히 예방하거나 관리하기 어렵다.

국가인권위원회가 2022. 5. 17.경 발표한 인공지능 개발과 활용에 관한 인권 가이드라인에서도 “인공지능 인권영향평가는 개발 및 출시 전에 실시하고 인공지능의 기능 또는 범위 변경시 평가를 갱신해야한다”고 권고하고 있다.

또한 앞서 살펴본 해외의 사례를 보더라도 인공지능기술에 대한 영향평가 수행 시기를 사전으로 한정하고 있지 않음을 알 수 있다. 캐나다 알고리즘영향평가의 경우 프로젝트 설계 초기에 우선 실시하도록 하고 있을 뿐 아니라 시스템의 생산 전에 두 번째로 실시하여 완화조치가 시스템에 제대로 반영되었는지 확인하도록 하는 절차를 두고 있다. 덴마크 <디지털활동 인권영향평가>의 경우 영향평가가 기업활동 전반에 대하여 반복적으로 시행되어야 한다고 하면서, 인권 위험 및 영향은 동일한 제품을 새롭거나 위험이 높은 시장에 출시할 때, 이용약관의 중대한 변경이 있을 때, 특정 시장에서 제품을 철수하는 등 디지털 사업, 제품 및 서비스의 규모, 범위, 사용 또는 적용이 변경될 때마다 재평가되어야 한다고 설명하고 있다.

---

128) Ad hoc Committee on Artificial Intelligence(2020).

따라서 인공지능 인권영향평가는 인공지능 시스템의 개발 전 계획 단계, 시행 전 단계에서 시행하는 것을 원칙으로 하되, 사전영향평가에만 한정하지 않고, 정기적, 사후적 평가를 병행하여야 할 것이다. 다만 모든 인공지능을 대상으로 여러 차례 영향평가를 시행하게 되면 막대한 인적, 물적 비용과 시간이 소요될 것이므로, 이를 고려하여 고위험 인공지능 중에서도 영향평가 결과 일정 기준 이상의 구체적인 위험이 확인된 경우만을 정기적, 사후적 평가 대상으로 삼는 등 적정 수준으로 제한할 필요가 있을 것이다.

다. 평가 주체

인공지능 인권영향평가의 수행단계에서 평가 업무를 인공지능 기술을 개발하거나 도입하여 사업에 활용하려는 주체로 하여금 스스로 수행하도록 할 것인지 그렇지 않고 별도의 독립된 조직 또는 제3의 기관이 수행하도록 할 것인지 문제되는데, 전자의 경우 기술 개발이나 도입과정에서 그 주체 스스로 일정한 기준을 가지고 평가하는 태도를 내면화할 수 있고, 타율적인 통제 없이도 영향평가의 목적에 부합하는 합리적인 결과물을 자율적으로 실현할 수 있도록 할 수 있다는 점에서 적지 않은 장점을 지닌다. 별도의 독립된 조직 또는 제3의 기관에 의뢰하는 경우 인공지능 기술이나 사업에 대한 정확한 정보나 깊이있는 이해가 결여되어 있을 가능성도 상당하다.

그러나 인공지능 기술의 개발자나 도입주체는 인권 영역이나 인권 침해 문제에 대한 전문성에는 한계가 존재할 수밖에 없다. 인공지능 기술의 개발자나 도입주체에 의한 자체평가의 경우 객관성, 중립성을 담보할 수 없고, 형식적인 절차로 운용할 가능성이 크다는 우려도 존재한다.

따라서 두 선택지 가운데 어느 한 경우를 택일하기보다는 각 선택지에 따른 부족한 점을 보완할 수 있는 방안을 적극적으로 모색해보아야 할 것으로 보인다.

예를 들어 인공지능기술의 개발 또는 도입 주체에 의한 자체평가 방식을 택하는 경우 형식적인 절차로 전락하지 않고 객관성과 중립성을 확보할 수 있도록 평가 결과(보고서)를 공개하거나 제3의 기관을 통한 사후 점검 절차를 두는 방식으로 보완할 수 있을 것이다.

반대로 별도의 독립된 조직 또는 제3의 기관이 주도하는 평가 방식을 택하는 경우 기

술개발, 사업 추진 등 관련 부서가 영향평가 과정에 적극적으로 참여하여 기술과 관련한 정확한 정보나 사업에 대한 이해도를 제고할 수 있도록 보완조치를 마련하여야 할 것이다. 또 적절한 평가 수행과 왜곡의 방지를 위하여 인공지능 기술 자체에 대한 전문성을 갖춘 인력이 평가 주체에 포함되도록 할 필요가 있다.

반드시 외부 기관에 속하지 않더라도, 인공지능 기술의 개발 또는 도입을 담당하는 개발 부서 또는 관련된 사업 부서와는 독립된 조직이나 부서(예를 들어, 인공지능 윤리, 인권 경영, ESG 경영 등을 담당하는 부서를 중심으로 평가팀을 구성할 수 있다)로 하여금 영향평가를 수행하도록 하면, 평가의 객관성과 중립성을 어느 정도 확보할 수 있을 뿐 아니라 개발 또는 사업 담당 부서와의 협업을 효율적으로 수행할 수 있는 장점이 있으므로 이 역시 개별 선택지의 단점을 보완할 수 있는 유용한 방안이 될 것이다.

한편 제3의 기관으로 하여금 평가를 수행하도록 하는 경우 국가인권위원회가 인권 전문성, 독립성 등을 검증하여 평가 수행 자격을 인증하도록 하여, 평가수행에 적합한 기관인지 여부를 관리하도록 할 수도 있을 것이다.

라. 의무성 및 주무기관

영향평가 실시에 관한 절차적 구속력과 영향평가 결과에 대한 내용적 구속력을 어떻게 부여할지 문제되는데, 금지대상 인공지능의 경우 개발이나 도입의 목적 자체로 수용 불가한 수준의 인권 침해 위험이 예상되기 때문에 즉각적으로 개발과 도입을 중단하도록 권고해야할 것이다.

나아가 고위험 인공지능 시스템의 경우 시스템의 개발 계획 단계 및 시행 단계에서 뿐만 아니라 사후적, 정기적으로 영향평가를 의무적으로 시행하도록 하고, 인권영향평가 시행 결과에서 인권에 미치는 부정적인 영향이나 편향성, 위험성이 드러난 경우 이를 방지하거나 완화하기 위한 조치사항을 수립하여 적용하도록 하며, 그 내용을 공개하도록 하고, 방지하거나 완화 조치를 취하기 전에는 개발과 활용을 중단하도록 권고할 수 있을 것이다.

고위험 인공지능 시스템에 해당하지 않는 경우에는 인권영향평가를 의무적으로 실시하도록 하지는 않고, 주무기관의 재량에 따라 직권 지정을 통해 영향평가를 실시하도록

하거나, 이해관계인의 신청에 따라 별도의 검토과정을 거쳐 지정을 하도록 하는 방식, 나아가 인공지능시스템의 개발 또는 도입 주체의 자발적 요구에 따라 영향평가를 시행하도록 하는 방식을 취하여도 무방할 것으로 보인다.

이 때 의무화를 실현하는 과정에서 주된 역할을 하는 주관부서는 결국 국가인권위원회가 되어야 할 것으로 보인다. 과학기술정보통신부가 정보통신정책연구원의 협의와 대통령 직속 4차 산업혁명위원회 전체회의를 거쳐 2020. 12. 22. <인공지능윤리기준>을 발표하였고, 그 직후 2020. 12. 23. <인공지능 법, 제도, 규제, 정비 로드맵>, 2021. 5. 14. <신뢰할 수 있는 인공지능 실현전략>을 발표하는 등 인공지능과 관련하여 적극적으로 개입하려는 의지를 보이고 있으나, 산업 진흥 부처인 과학기술정통부 보다는 인권에 미치는 영향을 객관적으로 평가하고 조치할 수 있는 규제부서가 주무기관으로 적합하다고 할 수 있다.

국가인권위원회는 독립성, 전문성 및 진정사건 처리 경험을 보유하고 있다는 점에서 인공지능 기술의 인권 준수에 대한 독립적인 감독, 인권영향평가 등 관련 지침과 권고 역시 전문적, 효과적으로 수행할 수 있을 것으로 기대된다. 국가인권위원회는 이미 2018년경 '공공기관 인권경영 매뉴얼' 적용 권고를 통해 공공기관에 대한 인권영향평가가 시행되도록 한 바 있고, 최근인 2022. 7. 13. '인권경영보고 및 평가지침'의 적용을 권고하는 등 공공기관의 인권경영 강화를 위한 역할을 수행하고 있다. 국가인권위원회가 인공지능을 개발하거나 도입하려는 경우에 대해서도 인권보호의 관점에서 영향을 평가하고 문제점이나 부정적인 영향이 확인되는 경우 방지 또는 완화조치를 권고하며, 이행이 되기 전에는 개발 또는 도입을 중단할 것을 권고 조치할 수도 있을 것이다.

마. 공개 및 점검

영향평가가 실효성을 지니기 위해서는 객관성과 중립성의 확보가 중요하므로, 이를 도모하기 위하여 인공지능 인권영향평가결과를 공개하도록 하거나, 제3의 기관을 통한 사후 점검절차를 거치도록 하는 것이 바람직하다. 이렇게 하면 인공지능기술의 개발 또는 도입 주체에 의한 자체평가 방식을 따르더라도 영향평가절차가 부실하게 시행되거나 편향된 형태로 진행될 우려를 어느 정도 해소할 수 있다.

인권영향평가는 공통적으로 이해당사자의 참여를 중시하는데, 이해당사자의 실질적인 참여를 도모하기 위해서는 투명성은 당연한 전제이다. 이를 위해서도 영향평가의 공개는 필수적이라고 할 수 있다.

캐나다 알고리즘영향평가의 경우 자동화된 의사결정 시스템을 생산하려는 기관이 온라인에서 직접 수행하도록 하고 있으나, 영향평가 최종결과를 캐나다 정부 웹사이트 및 캐나다 재정위원회가 지정한 서비스에 일반접근이 가능한 형식으로 공개하도록 하고 있다(훈령 6.1.4). 덴마크 <디지털활동 인권영향평가>는 공공기관과 민간의 개발사업자 또는 구매 사업자가 자율적으로 실시하도록 하되 영향평가 결과가 적절히 공개되어야 한다고 밝히고 있다.

다만 영향평가결과를 일반에 공개하는 경우 지적재산권이나 영업비밀의 노출이나 침해의 문제가 제기될 우려가 있으므로 요약본만을 공개하고, 구체적인 평가결과서는 제3의 기관에 보고하는 형식을 취하거나, 영업 비밀에 해당하는 부분을 제외한 나머지 내용만을 제한적으로 공개하도록 하는 방식을 고려해볼 필요가 있다.

영향평가결과의 보고 또는 공개시에 경영평가에서 가점을 주는 등의 방식으로 간접적으로 그 이행을 유인하거나, 사후에라도 영향평가수행과정의 적절성을 확인할 수 있도록 제3의 기관에 영향평가결과서를 보관하도록 하는 형태도 고려해봄직하다.

인공지능 기술의 개발 또는 도입을 담당하는 개발부서 또는 관련된 사업부서와는 독립된 조직이나 부서 또는 제3의 독립된 기관이 영향평가절차를 수행하도록 하는 경우 상대적으로 객관성과 중립성 확보에 용이한 측면이 있으나, 이 경우에도 평가수행의 적절성을 관리, 감독할 수 있도록 사후 점검절차를 두는 것이 바람직하다. 영향평가 결과서를 국가인권위원회에 제출하도록 하여 국가인권위원회가 영향평가 수행이 적절하게 이루어졌는지, 미흡한 점은 없는지 등을 사후 점검할 수 있도록 하면 영향평가의 실효성을 도모할 수 있을 것이다.[129] 인권영향평가 시행을 통해 확인된 구체적인 위험성의 정도에 따라 점검절차의 시행 여부나 방식은 차등 적용할 수도 있을 것이다.

129) 이는 개인정보보호영향평가와 유사한 방식이다. 개인정보영향평가의 경우 개인정보보호위원회가 지정하는 평가기관이 영향평가를 수행하고, 영향평가 결과를 개인정보보호위원회에 제출하도록 하며, 개인정보보호위원회가 영향평가결과에 대하여 의견을 제시할 수 있도록 하고 있다(개인정보 보호법 제33조 참조).

바. 다른 영향평가와의 관계설정

인권영향평가에서 인권의 개념은 포괄적인 특성을 지닌다. 나아가 인공지능 기술이 영향을 미칠 수 있는 기본권 또는 권리 역시 매우 다양하고 광범위하다. 그런데 우리나라의 영향평가제도는 다양한 영역에 대한 개별적인 영향평가 방식을 취하고 있어, 인공지능 인권영향평가를 통해 식별하고 확인하고자 하는 인권영향 중 일부가 이미 다른 영향평가를 통해 분석 또는 검토되었을 가능성이 존재한다.

특히 인공지능은 일반적으로 대규모 데이터셋의 패턴을 감지하여 작동하므로, 정보주체의 개인정보 침해의 가능성을 내재하고 있다. 이에 캐나다 알고리즘영향평가의 경우 평가 대상 인공지능 시스템이 개인정보와 관련된 경우 영향평가 질의에서 “개인정보보호 영향평가를 수행하거나 수행한 적이 있거나, 기존 영향평가를 갱신할 예정입니까”라고 묻도록 하였다.

네덜란드 FRAIA는 개인정보를 사용하는 인공지능의 경우 개인정보보호 영향평가는 FRAIA보다 더 좁은 범주의 평가라고 하면서, 질의 사항에 관한 논의에서 개인정보보호 영향평가에서 이미 수행된 분석을 포함시키는 것이 유용할 수 있다고 하였다.

우리나라의 경우에도 정보주체의 개인정보 침해가 우려되는 경우 위험요인의 분석과 개선사항 도출을 위한 목적으로 하는 개인정보영향평가제도가 이미 존재하므로 인권영향평가에서 이와 관련한 질의 항목의 경우 반복적으로 수행하지 않고 개인정보영향평가 절차의 수행 결과를 그대로 활용할 수 있을 것이다.

나아가 마찬가지로 인공지능 인권영향평가의 대상이 되는 인공지능기술을 활용한 사업 등에 관하여 환경영향평가가 시행되었다면 환경 영향에 대한 중복된 질의를 생략하고 이미 수행된 환경영향평가의 결과를 활용할 수 있을 것이다.

제2절 인공지능 인권영향평가도구(안)

1. 인공지능 인권영향평가 개요

본 인공지능 인권영향평가 제도의 주요 형식 및 절차는 다음과 같다.

가. 인공지능 인권영향평가의 대상 : 고위험 인공지능

법률에서 금지하는 인권 침해 또는 차별 대우를 목적으로 하거나 법률에서 금지하는 개인정보의 처리를 목적으로 하는 인공지능 등 위험의 완화 내지 제거가 불가능한 인공지능은 일응 ‘금지대상 인공지능’에 해당하여 인권영향평가의 대상에서 제외된다. 다만 이는 확정적인 기준일 수 없고, 하나의 예시로서 제안되는 것이며 향후 다양한 논의와 의견수렴을 통해 변동될 수 있다.

인공지능 인권영향평가가 대상인 인공지능은 공공기관이 직접 개발하거나 조달하는 모든 인공지능 및 민간에서 활용하는 인권 침해의 위험성이 높은 인공지능, 즉 고위험 인공지능이다. 물론 금지되는 인공지능의 기준과 마찬가지로 현 단계에서는 무엇이 고위험 인공지능인지에 대한 명확한 사회적인 합의가 존재하지 않으며, 고위험 인공지능에 대한 인권영향평가의 실시를 의무화하는 법률도 존재하지 않는다. 고위험 인공지능에 대한 인권영향평가 실시 의무화를 위해서는 고위험 인공지능의 범위 및 인권영향평가 수행 의무화를 내용으로 하는 입법화가 선행될 필요가 있다.

그러나 인권영향평가는 인공지능의 잠재적 위험을 체계적으로 검토하고 이해관계자와의 대화와 협력을 통해 사전에 방지, 완화하자는 취지로 시행된다. 따라서 국내외에서 고위험 인공지능이라고 거론되는 인공지능의 경우 본 인권영향평가를 수행할 것이 강하게 권고된다. 예를 들어, 적어도 아래 각 항목에 해당하는 인공지능은 일응 ‘고위험인공지능’에 해당한다고 보아 인공지능 인권영향평가의 대상으로 삼아야 할 것이다. 이는 유럽연합 인공지능법(안)에서 정한 고위험 인공지능 분류를 따르면서 군대 및 정보기관이 사용하는 인공지능의 경우도 추가한 형태이다. 다만, 이는 현 단계에서 예측되는 잠재적 위험을 기준으로 한 것으로 하나의 예시로서 제안되는 것이므로, 향후 위험성이 커

보이는 분야가 새롭게 확인되거나 드러나는 경우 항목이 추가될 수 있고, 반대의 경우 제외될 수 있다.

가. 항공, 자동차, 철도, 기계, 장난감의 안전 관련 구성요소이거나, 승강기, 무선 장비 및 의료 기기 등의 안전 관련 구성요소 또는 제품 그 자체인 경우
나. 사람의 생체정보를 활용하여 신원확인을 수행하는 경우
다. 교통, 수도, 가스, 전기 등 중요 사회기반시설의 관리·운용에 활용하는 경우
라. 소방, 응급의료 등 필수 공공·민간 서비스에 활용하는 경우
마. 채용, 인사평가 또는 직무 배치의 결정에 사용하는 경우
바. 공공 지원 혜택의 자격 및 수혜 적격성을 평가하기 위하여 사용하는 경우
사. 범죄의 수사, 공소의 제기 및 유지, 형 및 보안처분의 집행에 사용하는 경우
아. 이주, 망명 및 출입국 관리에 활용하는 경우
자. 사실의 인정 및 법률 해석, 적용 등 법관의 업무를 지원하는 데 사용하는 경우
차. 군 또는 정보기관에서 사용하는 경우

인공지능 인권영향평가가 고위험 인공지능을 대상으로 한다는 것이 고위험 인공지능이 아닌 경우에는 인권영향평가를 수행하는 의미가 없다는 뜻은 아니다. 설사 고위험 인공지능의 범위가 규정되더라도 인공지능의 위험성이 명확하게 단계화될 수 있는 것은 아니기 때문이다. 인공지능 인권영향평가의 취지가 인공지능의 잠재적 위험을 이해관계자와의 대화와 협력을 통해 사전에 파악하고 방지, 완화하자는 것이므로, 의무 대상에 이르지 않더라도 인권영향평가를 수행하는 것은 여전히 의미있고 권고할만한 일이다.

나. 인공지능 인권영향평가 시기

공공기관의 경우 위험도와 무관하게, 민간의 경우 고위험 인공지능을 개발하거나, 고위험인공지능을 사업 또는 정책의 기반기술로 도입하기 이전에 인권영향평가를 수행하되, 사전영향평가에만 한정하지 않고, 정기적, 사후적 평가를 통한 지속적인 관리와 모니터링을 전제한다. 고위험인공지능에 비하여 위험의 정도가 덜한 인공지능의 경우 국가인

권위원회의 직권 지정 또는 이해관계자의 요청에 따른 검토를 거쳐 국가인권위원회의 지정에 따라 인권영향평가를 수행할 수 있고, 개발 또는 도입 주체의 자발적인 요구에 의해서도 수행될 수 있다. 사전 영향평가의 시점은 인공지능기술 개발 또는 도입에 관한 구상이 구체화된 시점이다.

다. 인공지능 인권영향평가 수행 주체

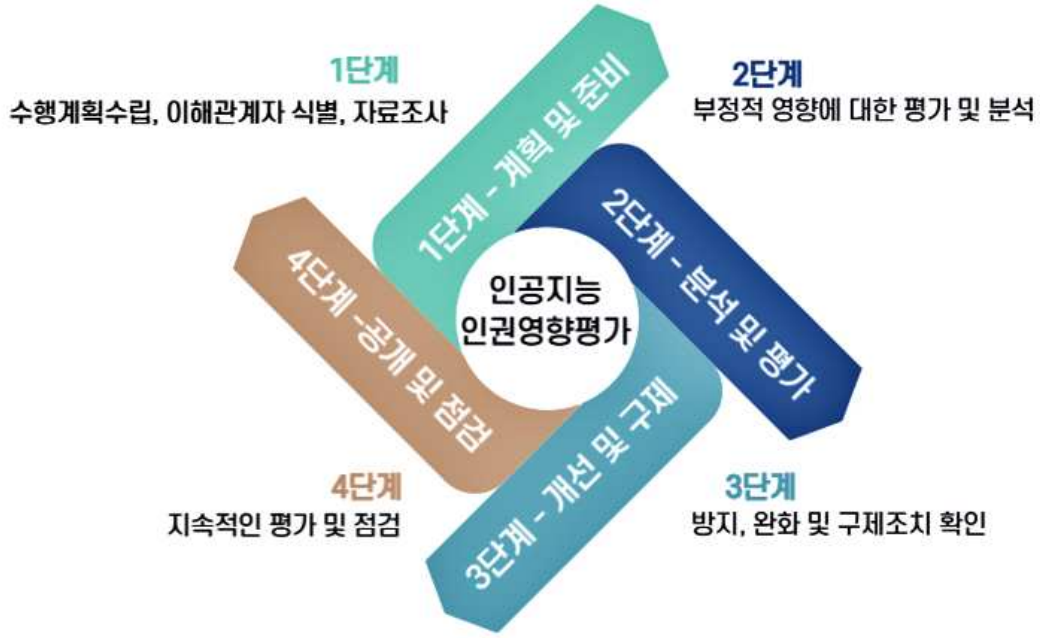
구체적인 영향평가의 수행은 인공지능의 개발 주체 및 관련된 사업부서와는 독립된 별도의 조직(예를 들어, 인공지능 윤리, 인권 경영, ESG 경영 등을 담당하는 부서) 또는 독립성과 인권 분야에 대한 전문성 및 인공지능 기술에 대한 전문성을 갖춘 제3의 기관이 수행하도록 한다.

라. 인공지능 인권영향평가의 절차

본 연구에서는 이상의 검토내용을 토대로 인공지능 인권영향평가의 이행단계를 아래와 같이 4단계로 범주화하였다.

- 1단계 : 계획 및 준비
- 2단계 : 분석 및 평가
- 3단계 : 개선 및 구제
- 4단계 : 공개 및 점검
- 공통 : 이해관계자의 참여

[그림 16] 인공지능 인권영향평가 이행단계



첫 번째 계획 및 준비 단계에서는 영향평가를 수행할 주체를 구성하고 관련한 기본적인 문헌을 조사하며, 평가 대상이 되는 인공지능 관련 사업 또는 정책을 선정하고 관련 이해당사자를 식별한다. 대상 기술 또는 사업과 관련한 기본 사실을 기입하면서 사업 계획 또는 진행 정도가 충분히 성숙되었는지 여부 등도 확인한다. 평가팀은 질의에 답하는 과정에서 인공지능과 관련한 정보를 수집하기 위하여 스스로의 판단하에 기획자, 개발자, 디자이너, 마케터 등 정보를 보유하고 있는 대상에게 질의하거나 의견을 구하고, 자료를 요청하여 확보할 수 있다.

영향 분석 및 평가 단계에서는 데이터, 알고리즘, 심각도의 각 분류 항목에 따른 체크리스트에 따라 그 영향의 정도를 평가한다. 이 단계는 인공지능 시스템에 의해 부정적인 영향을 받을 가능성이 있는 관련 권리를 식별하고 이에 대하여 영향의 정도까지 분석하는 과정을 포함한다. 개선 및 구제 단계에서는 방지, 완화 및 구제 조치가 이행되었는지를 확인한다.

공개 및 점검 단계에서는 완화조치들이 부정적 영향을 완화하였는지 지속적으로 평가하고 점검하는 단계이다. 투명성과 설명가능성을 충족하는지, 영향평가 결과를 공개하였는지 여부 등을 확인한다.

국내외 많은 영향평가에서 권고하고 있듯이, 이해관계자의 참여는 특정한 단계가 아니라 모든 단계에서 고려되거나 이행해야 할 사항이다. 예를 들어, 1단계에서는 인공지능 시스템의 성격이나 위험성에 따라 관련 이해관계자들을 식별하고 해당 이해관계자들과의 협의를 통해 정보를 수집하고 침해 가능성이 있는 인권을 식별하는 작업이 이루어져야 한다. 3단계(개선 및 구제)에서도 인권 침해를 방지, 완화하고 피해자의 권리를 구제하기 위한 여러 정책 제안에 대한 이해관계자의 의견을 수렴할 필요가 있다.

【참고】

• 덴마크 <디지털활동 인권영향평가>는 1단계 계획 및 범위 설정, 2단계 데이터 수집 및 맥락 분석, 3단계 영향 분석, 4단계 영향 예방, 완화 및 구제, 5단계 보고 및 평가로 구분하면서 모든 단계에서 공통적으로 권리 주체와 이해관계자의 참여가 기반이 되어야 한다고 설명하고 있다.

• 유럽평의회 역시 <인권・민주주의・법치 영향평가>에서 덴마크의 위 가이드를 인권영향평가 수행을 위해 참조할 수 있는 프레임워크로 제안하면서 유사한 평가 절차를 제안하고 있다.
- 1단계 : 인공지능에 의해 부정적 영향을 받을 수 있는 관련 권리의 식별
- 2단계 : 기술적/비기술적 요인을 포함한 해당 권리에의 영향평가
- 3단계 : 거버넌스 메커니즘 평가를 통한 위험 완화 요소 검토
- 4단계 : 지속적인 평가(evaluation)

• 영국 <NMIP 알고리즘영향평가>는 아래와 같은 7단계로 되어 있지만, 보건의료 분야, 그 중에서도 특정 데이터에 대한 제공 여부를 평가하기 위한 목적에 한정된 것이다. 다만, 여기서도 위험에 대한 평가, 이해관계자와의 협의 및 완화조치의 적용, 평가 결과의 공개 등의 요소를 포함하고 있음을 알 수 있다.
- 1단계 : 성찰적 수행
- 2단계 : 신청서 필터링
- 3단계 : 참여적 워크숍
- 4단계 : 종합 (성찰적 수행의 재검토)
- 5단계 : 데이터 접근 결정
- 6단계 : 출판
- 7단계 : 반복

• 캐나다 정부 <알고리즘영향평가> 도구는 자동화된 의사결정에 사용하는 인공지능 알고리즘을 도입하는 공공기관이 영향평가를 수행하는 것을 돕기 위한 것이다. 이 평가도구는 위험성과 완화 조치를 포함한 질문으로 구성되어 있으며, 평가 결과 나타난 위험성의 정도에 따라 추가적인 안전 조치를 취하도록 하고 있다. 이 역시 특정 부문을 위한 도구이기는 하지만, 위험평가, 완화조치, 평가 결과의 공개 등의 요소들을 포함하고 있다.

• 국가인권위원회 가이드라인에서도 “인공지능의 특성, 상황, 범위 및 목적”을 고려하여 “인권 가이드라인이 제시한 원칙 및 내용, 국제 인권 기준, 관련 법률에서 정한 의무 등”을 포함하도록 평가 기준을 제시하고 있으며, 인권영향평가를 통해 “인권침해 위험요인의 분석 및 개선 사항 등을 도출”하고 “인권에 미치는 부정적인 영향이나 편향성 및 위험성이 드러난 경우 이를 방지하거나 완화하기 위한 조치사항을 수립하여 적용”할 것을 권고하고 있다.

한편, 영향평가에 따른 결과서는 국가인권위원회에 제출되며, 국가인원위원회는 영향평가결과를 검토한 후 미흡한 점에 대한 개선을 권고하거나, 위험성에 대한 완화 조치 또는 제거 조치가 불가능하다고 판단하는 경우 개발 또는 활용의 중단을 권고하는 등 의견을 제시할 수 있도록 한다.

아래에서는 인권영향평가 과정에서 점검해야 할 항목을 체크리스트 방식으로 제시한다. 영향평가 수행자는 각 점검항목에 대해 “예 / 보완 필요 / 아니오 / 정보 없음 / 해당 없음” 중 하나를 선택할 수 있다. 여기서 ‘정보없음’은 질의에 답할 수 있는 관련 정보가 없어서 판단하기 힘들다는 의미이며, ‘해당없음’은 질의 자체가 평가 대상인 인공지능 시스템과 무관하다는 의미이다. 따라서 ‘해당없음’일 경우에는 해당 질의를 넘어가야 하지만, ‘정보없음’을 체크한 경우라면 향후 관련 정보를 찾기 위해 추가적인 노력을 할 필요가 있다는 의미이다. ‘정보없음’에 체크한 질의가 많을수록 인권영향에 대한 평가가 제대로 이루어지지 않을 가능성이 크다. ‘보완 필요’는 점검항목이 부분적으로만 충족된 경우를 의미한다. 점검항목의 개수를 최소화하는 과정에서 각 항목에 연관되지만 여러 개의 질의가 포함된 경우가 있을 수 있다. 이 중 일부의 질의만 충족한 경우에도 ‘보완 필요’를 체크하여 향후에 보완될 수 있도록 할 필요가 있는데, 서술식 공간을 활용하여 보완해야 할 내용을 기록할 것을 권장한다.

아래의 점검항목은 기업이나 기관을 규제하거나 세세한 의무를 부과하려는 것이 아니

다. 인공지능의 인권 침해 가능성을 탐지하고 완화하기 위해 고려할 사항이기는 하지만, 인공지능의 목적, 기술적 수단, 적용되는 분야 등은 매우 다양하고, 인권에 미치는 영향과 방법도 다를 수 있다. 따라서, 인권영향평가는 각 점검항목을 모두 충족해야만 하거나 질의에 대한 정답을 맞추는 것이라기 보다는 점검항목을 매개로 하여 이해관계자 사이의 소통과 토론, 그리고 서로의 역량을 강화하는 과정이 되어야 한다. 점검항목을 충족하지 못하더라도 합당한 이유나 다른 방안을 제시하는 방식도 있을 수 있다. 이에 모든 항목에 대해서 체크리스트 방식의 표기와 함께, 관련된 구체적인 정보를 서술식으로 설명하는 공간을 제공하고 있다. 이를 활용하여 기술에 전문성이 없는 이해관계자라도 이해할 수 있도록 가능한 쉽고 상세한 설명을 제공할 것을 권장한다.

2. 인공지능 인권영향평가도구

【1단계 : 계획 및 준비】

1단계에서는 인공지능 인권영향평가를 수행할 팀(이하 평가팀)이 구성되어, 인권영향평가를 제대로 수행할 수 있는 여건을 점검하고 수행 계획을 세우며, 이해관계자를 파악하고, 관련 자료를 조사·수집하게 된다. 아래의 질의는 인권영향평가를 수행하는 팀이 검토해야 할 사항을 다루고 있으며, 평가팀은 이 질의에 답하기 위해 인공지능 시스템의 개발자(혹은 개발업체)를 비롯하여, 해당 인공지능 시스템의 사용자 및 영향을 받는 이용자 등 이해관계자로부터 자료를 수집하거나 의견을 청취할 수 있다. 이를 위해 평가팀은 아래 질의를 참고하여 해당 인공지능 시스템의 맥락에 적합하게 수정하여, 개발자를 포함한 각 이해관계자에게 자료 제공이나 의견을 요청하기 위한 질의 문서를 작성할 수 있을 것이다. (참고로 개발자 혹은 개발업체에 인공지능 시스템 관련 정보의 제공을 요구할 사항의 예시를 본 도구의 뒤에 부록으로 정리하였다.) 필요할 경우, 평가팀은 인터넷을 통해 관련 자료를 검색, 수집할 수도 있다. 이렇게 수집된 자료 및 이해관계자의 의견을 참고하여 2단계에서 인공지능 시스템이 인권에 미치는 부정적 영향에 대해 분석 및 평가를 하게 된다.

가. 인권영향평가 계획

Q1-1-1. 인권영향평가의 대상이 되는 인공지능 시스템 혹은 프로젝트는 무엇입니까. 인공지능 시스템 혹은 프로젝트명 :

종종 여러 시스템이나 프로젝트가 혼재되어 있을 수 있다. 우선 본 인권영향평가의 대상이 되는 인공지능 시스템, 혹은 개발 프로젝트를 명확하게 규정한다.

Q1-1-2. 인권영향평가를 수행하는 책임자는 누구입니까. 책임자의 성명과 소속 :

Q1-1-3. 평가팀은 인권영향평가를 수행하기에 충분한, 인공지능 기술 및 인권에 대한 전문성을 갖추고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

평가팀의 구성, 팀원의 역할, 전문분야 등을 설명하시오.

설명 ( )

인권영향평가의 책임성을 위해 책임자의 성명과 소속을 기록한다. 평가팀은 인공지능의 개발 주체 및 관련된 사업부서와는 독립된 조직 내부의 별도의 조직(예를 들어, 인공지능 윤리, 인권 경영, ESG 경영 등을 담당하는 부서를 중심으로 평가팀을 구성할 수 있다) 혹은 독립성과 인공지능 및 인권 분야에 대한 전문성을 갖춘 제3의 기관이 될 수도 있다. 평가팀은 인권 뿐만 아니라 인공지능에 대한 전문성도 갖춰야 한다. 그렇지 않으면, 형식적이거나 가식적인 대응을 제대로 식별할 수 없기 때문이다. 신뢰성있는 인권영향평가를 위해 평가팀의 역량 및 구성 등을 점검하도록 한다.

Q1-1-4. 조직 내에 인권영향평가 수행의 요건, 주체, 절차 등을 상세히 규정한 정책을 두고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q1-1-5. 인권영향평가를 내실있게 수행하는데 충분한 인적, 재정적 자원이 확보되어 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q1-1-6. 인권영향평가 결과의 수용 여부를 결정할 수 있는 조직 내 최종 책임자 혹은 책임단위에 인권영향평가 결과보고서를 보고하는 절차가 명확하게 규정되어 있습

니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q1-1-7. 인권영향평가를 내실있게 수행할 수 있도록, 평가팀이 평가 대상이 되는 인공지능 시스템의 개발 혹은 활용과 관련한 부서 및 담당자에게 협조를 요청하고, 인권영향평가에 필요한 핵심 자료에 접근할 수 있는 권한이 보장되어 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

우선 인공지능 인권영향평가 수행을 위한 계획이 수립되어야 한다. 이때 임시방편적인 인권영향평가가 되지 않기 위해서는 인공지능 인권영향평가를 위한 체계가 조직 내 정책으로 이미 수립되어있는 것이 바람직하다. 조직 내에서 인권영향평가 수행의 요건, 주체, 절차 등을 마련해놓고 있다면 영향평가를 수행해야 할 상황이 발생했을 때 보다 용이하게 평가 계획을 수립할 수 있다. 만일 인공지능 인권영향평가를 처음 수행하는 것이라면, 향후 인권영향평가 자체에 대한 평가와 검토(본 인권영향평가에서는 제4단계에서 인권영향평가 자체에 대한 평가와 검토를 하도록 하고 있다)를 통해 조직 내 정책으로 통합하는 것이 바람직하다. 따라서 인권영향평가를 수행하는 팀은 결과 보고서에 인권영향평가의 조직 내 정책화에 대한 권고사항도 포함할 필요가 있다.

인권영향평가 수행을 위해 충분한 인적, 재정적 자원이 확보되어 있지 않다면 인권영향평가 절차는 오히려 고위험 인공지능 도입을 정당화하는 요식행위가 될 수 있다. 따라서 인적, 재정적 자원이 확보되어 있는지 사전에 점검할 필요가 있다. 또한 인권영향평가가 의미가 있으려면 실제 평가의 결과가 후속 조치에 반영될 수 있도록 의사결정 단위에 전달되어야 한다.

인권영향평가의 객관적 수행을 위해서 인공지능의 개발 주체 및 관련된 사업부서 외부의 평가팀(조직 내부의 독립 부서 혹은 제3의 기관)이 평가를 수행하는 만큼, 평가 대상에 대한 정확한 이해를 위해서는 담당 부서나 담당자(예를 들어, 인공지능 시스템의

개발자나 발주자 등)의 협조가 필수적이다. 인권영향평가를 수행하기 이전에 관련 부서나 담당자의 협조가 가능한지 여부, 그리고 관련 자료에 대한 접근이 어느 정도 가능한지 여부에 대해 확인할 필요가 있다. 담당자 협조와 자료 접근이 보장되지 않는다면 인권영향평가의 원활한 수행이 어려울 수 있다.

인권영향평가 수행을 위한 자료에 영업비밀이 포함되어 있을 수 있다. 그러나 영업비밀을 근거로 평가팀이 관련 자료에 접근할 수 없다면 인권영향평가를 수행하는 의미가 없어질 것이다. 다만, 인권영향평가 수행 과정에서 영업비밀 침해가 발생하지 않도록 비밀서약 등 적절한 안전조치를 마련할 필요는 있다. 또한 인권영향평가의 결과를 공개할 때에도 영업비밀인지 여부를 고려해야할 것이다.

Q1-1-8. 인공지능 시스템은 어떠한 문제를 해결하기 위한 것입니까, 즉 인공지능 시스템이 달성하고자 하는 목적 및 의도된 용도는 무엇입니까.
인공지능 시스템의 목적 :

인권영향평가의 대상이 되는 인공지능 시스템에 대한 설명을 제공한다. 해당 인공지능 시스템이 어떠한 문제를 해결하고자 하는 것인지, 의도된 용도가 무엇인지 명확히 하는 것이 중요하다. 그래야 굳이 해당 인공지능 시스템이 필요한 것인지, 다른 대안적인 방법은 없는 것인지, 인공지능이 야기할 수 있는 인권침해 위험성에 비해 인공지능이 창출하는 가치가 균형적인지 등에 대한 판단이 가능하기 때문이다. 또한, 대상이 되는 인공지능 시스템에 대한 설명은 관련 이해관계자에게 제공되고 향후 일반에 공개될 수 있다. 따라서 기술 전문가가 아닌 사람도 이해할 수 있도록 평이한 용어로 작성하는 것이 바람직하다.

기존의 인공지능 시스템을 특정 서비스에 적용하는 경우도 있을 수 있다. 이때 인공지능 시스템은 해당 서비스를 위한 원천기술이 된다. 특정 서비스에의 적용에 대해 인권영향평가를 하는 경우, 해당 인공지능 시스템을 활용한 특정 서비스가 해결하고자 하는 목적을 작성하면 된다.

【참고】

• 네덜란드 <기본권 알고리즘영향평가>는 “1부 : 왜 하는가”에서 알고리즘을 도입하려는 이유, 해결하고자 하는 문제, 알고리즘의 목적, 추구하는 공공 가치, 알고리즘 사용의 법적 근거, 이해관계자, 알고리즘의 개발 및 사용에 대한 책임 할당 등을 검토하도록 하고 있다.

• 덴마크 <디지털활동 인권영향평가>는 인권영향평가의 거버넌스 구조가 명확하게 설명되어 있는지, 보고서(전체 또는 일부)의 발간을 비롯한 보고 요구사항이 명확히 규정되어 있는지, 고위 경영진/임원의 역할 및 관련 후속 활동의 내부 소유구조가 명확하게 설명되어 있는지, 예산이 명확하고 지정된 평가를 수행하기에 충분한지, 인권영향평가의 기간이 특정되어 있으며 평가 완수에 필요한 연구와 이해 관계자 참여에 충분한 시간을 보장되는지 검토하도록 한다.

• 영국 <NMIP 알고리즘영향평가>는 1단계에서 프로젝트에 대한 배경 정보를 기록하도록 하고 있다. 이는 평가도구의 후반부에서 윤리적 고려 사항 및 잠재적인 피해를 검토하는데 도움을 주기 때문이다. 프로젝트의 목적, 의도된 용도, 조직에 대한 설명 및 유형, 시스템의 입력 및 출력, 시스템의 영향을 받는 이해관계자 등이 이에 포함된다. 프로젝트의 목적은 독자가 기술 지식이 많지 않다고 가정하고 낯선 사람에게 설명하듯이 정리할 것을 요구한다.

• 캐나다 정부 <알고리즘영향평가> 도구에서는 제1장에서 답변자의 이름 및 직책, 기관명과 부서명, 프로젝트명, 프로젝트의 단계(설계단계 혹은 구현단계), 프로젝트에 대한 설명 등 프로젝트 세부정보를 작성하도록 하고 있다. 제2장에서는 의사결정 과정에 자동화를 도입하는 동기가 무엇인지 체크하도록 하고 있고, 제3장은 프로젝트의 위험성을 설명하도록 한다.

• 금융위원회 <금융분야 AI 개발·활용 안내서>는 “금융회사 등은 AI 시스템의 전 과정에 걸쳐 AI 활용에 따라 나타날 수 있는 잠재적 위험을 인식·평가하고, 이를 관리·최소화하는 방안을 검토하는 등 AI 활용으로 인한 잠재적 위험을 관리하는데 필요한 위험관리정책을 마련해야 한다”는 원칙을 제안하고 있다.

Q1-1-9. 해당 인공지능 시스템이 적용되는 분야에서 인공지능 시스템의 기능, 요건, 제한 등에 영향을 미치는, 인권 보호를 위해 요구하고 있는 법령상의 요건(법률, 시행령, 시행규칙 등의 관련 조항)이 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

만일 있다면, 해당 법률, 시행령, 시행규칙 등의 관련 조항은 무엇입니까. 설명 ( )

특히 공공기관에서 특정 인공지능 시스템을 도입할 경우 법적 근거가 필요하다. 법에서 인공지능 시스템의 기능, 요구조건, 제한 등을 규정하고 있을 경우 이를 준수할 필요가 있으며, 이는 인권영향평가 과정에서도 고려되어야 한다. 예를 들어, <공기업·준정부기관의 인사운영에 관한 지침>은 인사를 "공정하고 투명하게 운영"할 것을 원칙으로 하고 있으며 세부적인 규정을 두고 있다. 이는 인공지능 시스템의 지원을 받아 인사운영을 할 때에도 적용되어야 하므로, 이를 위한 인공지능 시스템을 개발할 때부터 고려되어야 한다.

공공기관이 민간업체가 개발한 인공지능 시스템을 조달하는 경우가 많기 때문에, 인공지능 시스템을 개발하는 민간업체 역시 법적 근거의 존재 여부와 그것이 인공지능 시스템의 기능, 사양, 활용 범위 등에 미칠 영향에 대해 고려할 필요가 있다. 만일 인공지능 시스템이 여러 국가에 걸쳐서 활용된다면 각 국가마다의 법적 근거를 파악할 필요가 있다. 새로운 국가나 지역에서 활용될 경우, 해당 국가나 지역의 맥락에서 특별하게 인권 침해의 우려가 제기된다면, 새롭게 인권영향평가를 수행해야 할 수도 있다.

Q1-1-10. 인공지능 시스템이 인권에 미치는 영향을 평가하기 위하여 조직 내외부의 다양한 이해관계자의 의견을 검토할 필요가 있습니다. 다양한 이해관계자를 인권영향평가 과정에 참여시키기 위해서는 우선 누가 이해관계자인지 파악해야 합니다. 아래 질의에서 이해관계자가 누구인지 가능한 구체적으로 적어주세요.
Q1-1-10-1. 해당 인공지능 시스템에 대한 공정한 인권영향평가를 위해, 조직 내부에서 해당 인공지능 시스템의 개발 및 운영에 관련된 이해관계자(예를 들어, 기획, 개발, 디자인, 유지보수, 정책, 데이터 거버넌스, 영업 등 담당 부서)의 참여가 중요합니다. 이를 위해 조직 내부에서 참여할 수 있는 이해관계자는 누구입니까. 설명 ( )

Q1-1-10-2. 해당 인공지능 시스템에 대한 공정한 인권영향평가를 위해, 조직 외부에서 해당 인공지능 시스템의 개발 및 운영에 관련된 이해관계자(예를 들어, 외부 개발업체, 위탁업체, 유지보수업체, 감독기구, 전문가 집단 등)의 참여 역시 중요합니다. 이를 위해 조직 외부에서 참여할 수 있는 이해관계자는 누구입니까.

설명 ( )

Q1-1-10-3. 인공지능 시스템의 사용자는 누구입니까.

설명 ( )

Q1-1-10-4. 인공지능 시스템의 사용으로 영향을 받는 사람이나 집단은 누구입니까.

설명 ( )

Q1-1-10-5. 인공지능 시스템의 사용으로 영향을 받는 개인이나 집단에 아동, 노인, 장애인, 여성, 외국인, 성소수자, 저학력자, 비정규직 노동자, 경제적 약자, 낙후지역 등 취약하거나 소외된 집단이 포함되어 있다면 구체적으로 적어주세요.

설명 ( )

다른 영향평가와 달리 인권영향평가의 경우 이해관계자와의 소통과 협의가 특히 강조된다. 예를 들어 개인정보 영향평가는 이해관계자인 정보주체와의 협의를 의무 사항으로 두고 있지는 않다. 반면, 인권영향평가는 이해관계자, 특히 자신의 인권이 침해될 가능성이 있는 피해당사자의 참여를 핵심 원칙으로 두고 있다. 본 영향평가안에서도 “이해관계자의 참여”를 공통 절차로 두고 있는데 그 출발점이 되는 것이 해당 인공지능 시스템 관련 이해관계자에 대한 파악이다. 즉, 여기서 이해관계자를 파악하고자 하는 취지는 인권영향평가 과정에서 이해관계자의 다양한 관점과 의견을 반영하기 위한 것이다.

인공지능의 개발자와 사용자는 다른 경우가 많다. 예를 들어, 보건의료 분야의 인공지능의 경우, 개발자는 개발업체의 개발팀이 될 수 있겠지만 사용자는 의사나 간호사가 될

것이며, 영향을 받는 주체는 환자가 될 것이다. 장애인 지원 서비스 관련 인공지능의 경우 개발은 민간업체가 하더라도 해당 서비스를 제공하는 공공기관이 사용자가 될 수 있으며, 장애인이나 활동 보조인이 영향을 받는 이해관계자가 될 수 있다.

인공지능 시스템 혹은 서비스의 개발자 역시 다양한 역할 및 층위에 있을 수 있다. 예를 들어, 인공지능 모델 아키텍처의 설계자·구현자, 인공지능 모델 훈련(학습)을 위한 개발자, 훈련(학습) 완료된 인공지능 모델이 추론·예측·생성을 실행할 수 있도록 하는 개발자, 인공지능 모델과 사용자 간의 인터페이스(API) 개발자, 인터페이스에 기반하여 최종사용자가 사용할 수 있는 서비스 혹은 솔루션을 만드는 개발자 등이 있을 수 있다. 어떠한 개발자를 이해관계자로 포함할 것인지는 인권영향평가의 대상이 되는 인공지능 시스템이나 서비스의 맥락에 따라 달라질 것이다. 예를 들어, 타 업체의 인공지능 시스템 API를 사용하여 특정 서비스에 적용하는 기업이라면, 인공지능 시스템 API 개발업체와 자사의 서비스 개발자를 이해관계자로 포함할 수 있을 것이다. 즉, 어떠한 이해관계자를 본 인권영향평가를 위해 고려할 필요가 있는지에 대한 관점에서 판단한다.

영향을 받는 이해관계자가 취약계층이나 소수자 집단일 경우 자신의 정당한 권리를 주장하기 힘든 위치에 있을 수 있다. 이 경우 장애인 권리 옹호 인권단체가 이해관계자로 참여할 수도 있다.

【참고】

• 네덜란드 <기본권 알고리즘영향평가>는 “1부 : 왜 하는가”에서 알고리즘 사용의 법적 근거를 검토하도록 하고 있다.

• 미 의회 <2022년 알고리즘 책무성법(안)>에서는 영향평가를 수행하는 주체가 관련 내부 이해관계자(피고용인, 윤리팀, 담당기술팀 등) 및 독립적인 외부 이해관계자(영향을 받는 집단의 대표자 혹은 옹호자, 시민사회, 기술 전문가 등)와 가능한 자주, 의미있는 협의(참여 설계, 독립적인 감사, 의견 수렴 등)를 하도록 요구하고 있다(sec.3.(b)(G)).

• 영국 <NMIP 알고리즘영향평가>는 알고리즘 시스템의 영향을 받는 이해관계자를 파악할 때 가능한 구체적으로 (예를 들어 임상의, 간호사, 병원 행정 직원, 특정 유형의 환자 등) 작성할 것을 권고한다(1.d).

• 덴마크 <디지털 업무 인권영향평가>는 인권영향평가에서 고려해야 하는 관련 배경 정보(예) 다른 인권실사 활동, 개인정보 영향평가, 윤리적 영향평가 및 기타 평가를 실시한 결과 등)가 과제 서술에 포함되어 있는지 질의한다.

• 유엔 <기업과 인권 이행지침>은 기업이 인권에 미칠 수 있는 부정적 영향을 평가할 때, 내부 또는 외부의 독립적인 인권 전문가를 활용하고 그리고 잠재적으로 영향을 받을 수 있는 집단과 기타 관련 이해관계자 대상의 실질적 자문을 포함하도록 하고 있다.

나. 조사

Q1-2-1. 인공지능 시스템이 인권에 미치는 영향을 이해하기 위해서는 해당 시스템에 대한 이해가 필요합니다. 데이터셋, 알고리즘 등 해당 인공지능 시스템과 관련된 정보(예를 들어, 데이터셋이나 알고리즘 등의 특성 및 이에 대한 평가, 외부업체의 제품을 구매할 경우 관련한 설명서, 사전학습 모델 가중치 등)를 확보하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

인권영향평가를 위해서는 평가 대상 및 환경에 대한 충분한 자료를 확보해야 할 것이다. 여기에는 평가 대상인 인공지능 시스템에 대한 정보, 특히 학습 및 테스트에 사용된 데이터셋과 알고리즘, 사전학습 모델 가중치 등에 대한 정보가 포함된다. 이들 데이터셋과 알고리즘 등을 모두 개발업체 스스로 만들지 않고 외부의 소스로부터 가져올 수도 있다. 다른 기업이 만든 인공지능 시스템 API를 활용하여 특정 서비스를 개발할 수도 있다. 만일 민간기업이 개발한 인공지능 시스템을 사용하는 공공기관이 인권영향평가를 수행한다면, 해당 민간기업에 관련 자료를 요구해야 할 필요가 있는데, 이때 해당 데이터셋, 알고리즘 등과 관련된 정보를 여기 1단계에서 파악한다. 예를 들어, 해당 데이터셋이나 알고리즘에 대한 학계 및 기술계의 평가가 있다면 이를 수집할 수도 있고, 외부업체가 개발한 제품을 사용할 경우 관련한 설명서를 요청할 수도 있다. 이에 대한 평가와 분

석은 2단계에서 수행된다.

Q1-2-2. 인공지능 시스템이 도입, 활용될 분야 혹은 시공간적인 특성 및 맥락과 관련된, 인권에 영향을 미칠 수 있는 요소에 대한 자료를 확보하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

인공지능과 같은 특정한 기술이나 서비스가 인권에 미치는 영향은 단지 기술이나 서비스의 특성에만 달려있는 것은 아니며, 해당 기술이나 서비스가 도입, 활용되는 사회적, 지역적, 역사적 맥락에 따라 달라질 수 있다. 예를 들어 2018년 페이스북은 미얀마에서 인권영향평가를 수행한 바 있는데, 미얀마 페이스북을 통해 혐오 발언이 확대되고 오프라인에서의 폭력과 학살로 이어진 것에 대한 비판이 제기되었기 때문이다. 이처럼 똑같은 페이스북 서비스라도 국가와 지역에 따라 인권에 미치는 영향이 다르게 나타날 수 있다. 얼굴인식 대량감시를 용이하게 하는 인공지능의 경우 이것이 활용되는 지역의 법제, 정부의 특성, 문화 등에 따라 그 영향이 크게 달라질 것이라는 점은 쉽게 짐작할 수 있다. 따라서 인권영향평가의 대상이 되는 인공지능 시스템 혹은 서비스가 실제 활용되는 시공간적인 특성 및 맥락에 관련된 정보를 파악할 필요가 있다.

Q1-2-3. 앞서 파악한, 인공지능 시스템의 이해관계자로부터 해당 시스템이 인권에 미칠 영향에 대한 의견을 수렴하거나 협의하고 이를 문서화 하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q1-2-4. 이해관계자 의견을 수렴하거나 협의할 때 다음과 같은 내용을 포함합니까.

- 협의한 이해관계자의 성명, 소속, 연락처
- 협의한 일자
- 인공지능 시스템에 대해 이해관계자에게 제공한 자료
- 인공지능 시스템에 대한 이해관계자의 의견

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-2-5. 인공지능 시스템의 활용으로부터 영향을 받는 이해관계자, 특히 취약하거나 소외된 집단과의 협의를 포함하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-2-6. 관련 자료를 수집하거나 이해관계자의 의견을 수렴할 때 자료의 기밀성을 유지하고 이해관계자의 개인정보를 보호할 수 있는 조치를 취하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

이해관계자와의 의견수렴 및 협의 방식은 이해관계자의 성격, 규모, 구체적인 맥락에 따라 달라질 수 있다. 원격으로 소통할 수도 있고 오프라인 회의를 할 수도 있다. 이해관계자 별로 따로 소통할 수도 있고 여러 이해관계자가 함께 모여서 토론하는 방식일 수도 있다. 예를 들어, 영국 <NMIP 알고리즘영향평가>의 경우, 다양한 이해관계자들과 참여 워크숍(participatory workshop)을 수행하도록 하고 있다. 이러한 과정을 통해 이해관계자는 역량을 강화하거나 개별적인 소통에서는 가능하지 않았던 깨달음을 얻을 수 있다. 그러나 과거 전통적인 인권영향평가와 달리, 인공지능 시스템의 경우 글로벌한 이용자를 대상으로 할 수도 있으며, 이럴경우 특정 지역의 이용자만을 대상으로 협의하는 것이 적절하지 않을 수 있다. 직접적으로 관련된 이해관계자뿐만 아니라, 인권단체, 감독기구, 학술연구자 등과의 협의도 고려할 수 있다.

이해관계자와의 협의는 가능한 다양하고 포괄적으로 이루어지는 것이 바람직하다. 인공지능 시스템의 활용으로부터 영향을 받는 이해관계자, 특히 취약하거나 소외된 집단이 있을 경우 반드시 포함할 필요가 있다.

수집된 데이터는 기밀로 유지할 필요가 있으며, 이해관계자의 개인정보가 필요 이상으

로 수집, 처리, 공개되지 않도록 주의해야 한다. 특히 취약하거나 소외된 집단 등 어떤 이해관계자들은 인권영향평가에 참여할 경우 보복의 위협에 직면할 수 있기 때문에, 적절한 보호 조치가 없다면 이해관계자의 참여에 제약이 발생할 수 있다.

【참고】

• 미 의회 <2022년 알고리즘 책무성법(안)>에서는 협의한 이해관계자의 연락처, 협의 일자, 협의의 조건 및 절차에 대한 정보(이해관계자와 대상 기업 간 법적 또는 재정적 합의의 존재 유무 및 성격, 이해관계자와 상호교류한 모든 데이터, 시스템, 설계, 시나리오 및 기타 문서와 자료, 자동화된 의사결정 시스템 또는 증강된 중요 의사결정 프로세스의 개발 또는 배치를 변경하는 데 사용된 이해관계자의 모든 권장사항, 사용되지 않은 권장사항 및 미사용 근거)들을 파악하고 설명하도록 하고 있다.

• 캐나다 정부 <알고리즘영향평가> 도구에서는 제10장에서 내부 이해관계자(전략 정책 및 계획 부서, 데이터 거버넌스 부서, 프로그램 정책 부서 등) 및 외부 이해관계자(시민사회, 학계, 산업계 등)의 자문을 받았는지 여부에 대해 질의한다.

• 영국 <NMIP 알고리즘영향평가>는 수행자가 시스템의 영향을 받는 다양한 이해관계자들과 참여 워크숍(participatory workshop)을 수행하면서 해당 프로젝트의 잠재적인 영향을 논의하도록 한다. 워크숍은 연령, 성별, 지역, 민족적 배경, 사회경제적 배경, 건강 상태 또는 치료 접근성에 걸쳐 알고리즘의 영향을 받을 수 있는 인구의 다양성을 반영하는 패널 8-12명으로 구성된다. 신청자(평가 수행자) 역시 워크숍에 참여하여 프로젝트에 대해 설명하고 패널의 질문에 답변하며 토론을 경청한다.

• 덴마크 <디지털활동 인권영향평가>는 수집된 데이터는 기밀로 유지되어야 하며 연구자는 연구자가 게시하거나 다른 방식으로 사용하는 데이터에서 개별 참가자를 식별할 수 없도록 요구한다. 이는 문제가 민감하고 참가자가 보복 위험에 직면할 수 있는 인권영향평가의 경우 특히 중요하다. 따라서 연구자는 참여자의 개인정보와 답변에 대해 강력한 데이터 보호 조치를 취해야 한다.

【2단계 : 분석 및 평가】

2단계, 영향 분석 및 평가는 크게 두 부분으로 구성된다. 첫째는 인공지능 기술과 관련된 것이며, 둘째는 인권 자체에 미치는 영향과 심각도에 초점을 맞춘다. 인공지능 기술과 관련된 영향 분석 및 평가는 해당 인공지능 시스템이 사용하고 있는 특정한 데이터 및 알고리즘에 대한 지식을 바탕으로 하기 때문에, 인권영향평가 수행자는 이와 관련된 정보에 접근할 수 있어야 하며, 개발자 혹은 외부 개발업체의 지원과 협력을 필요로 할 수 있다.

가. 인공지능 기술과 관련된 영향 분석 및 평가

(1) 개인정보보호

Q2-1-1. 인공지능 시스템이 개인정보보호위원회 <인공지능(AI) 개인정보보호 자율점검표>의 모든 의무/권장 조항을 준수하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-2. 해당 인공지능 시스템의 개발 혹은 운영 과정의 개인정보 처리가 개인정보보호법 상 개인정보 영향평가를 의무적으로 수행해야 하는 경우, 개인정보 영향평가를 수행하였는지 확인하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

인공지능 시스템 개발 및 운영 과정에서 수집, 처리되는 개인정보는 당연히 개인정보보호법에 따라 적법하게 처리되어야 한다. 개인정보 보호법의 준수 여부를 확인하기 위해 데이터 최소화의 원칙, 개인정보 처리의 적법 근거, 정보주체의 권리 보장 여부, 개인정보에 대한 안전성 조치 등을 검토할 필요가 있다. 개인정보보호위원회가 이미 발표한

<인공지능(AI) 개인정보보호 자율점검표>가 있으므로 본 영향평가안에서 세부적인 질의를 포함하는 것보다 이 자율점검표의 요구사항을 준수하고 있는지 확인하도록 하였다. 한국의 개인정보 보호법은 (2022년 10월 현재) 프로파일링 및 자동화된 결정과 관련한 규정 등 인공지능 기술 환경을 반영한 규범을 아직 포함하고 있지 않아 일정한 한계가 있지만, 인공지능 시스템이 한국의 개인정보 보호법을 준수하고 있는지에 대한 점검을 위해서 자율점검표를 활용하는 것이 바람직하다.

정보주체의 권리와 자유에 중대한 위험을 야기하는 개인정보 처리가 수반될 경우 법에 의해 개인정보 영향평가를 의무적으로 수행해야 할 수도 있다. 한국의 경우 공공기관이 일정 기준 이상의 개인정보 처리가 수반되는 개인정보 파일이나 시스템을 운용할 경우 의무적으로 개인정보 영향평가를 수행해야 하며, 유럽연합 일반개인정보보호규정(GDPR)은 공공 및 민간영역 모두에서 개인정보 영향평가를 수행하도록 하고 있다. 유럽 시민의 개인정보 처리를 예정하고 있다면 한국의 인공지능 개발자 역시 GDPR에 따른 개인정보 영향평가를 수행해야 할 수 있다.

해당 인공지능 시스템이 개인정보 보호법에 따른 개인정보 영향평가를 의무적으로 수행해야 하는 경우가 아니더라도, 본 인공지능 인권영향평가를 자율적으로 수행할 수 있다. 본 인권영향평가안에서는 개인정보에 미치는 영향을 세세하게 다루지는 않으며, 개인정보자기결정권에 미치는 영향에 대한 검토는 <인공지능(AI) 개인정보보호 자율점검표> 혹은 개인정보 영향평가 절차에 위임하고 있다. 개인정보 영향평가를 이미 수행했다면 그 내용을 본 인권영향평가에 반영할 수 있다. 혹은 개인정보 영향평가 의무 대상임에도 불구하고 아직 수행하지 않았다면, 개인정보 영향평가를 적절한 시점에 수행하도록 권고할 수 있으며, 그때 본 인권영향평가의 관련 내용을 고려하도록 할 수 있다.

【참고】

• 과학기술정보통신부 <2022 인공지능 윤리기준 실천을 위한 자율점검표(안)>은 10대 핵심요건 중 하나로 “프라이버시 보호”를 두고 있고, 이를 점검하기 위해 개인정보보호위원회 <인공지능(AI) 개인정보보호 자율점검표>에 따른 점검을 수행하였는지 확인하고 있다.

• 캐나다 정부 <알고리즘영향평가> 도구에서는 시스템에 개인정보 사용이 포함된 경우 개인정보 영향평가를 수행하였는지, 프로젝트의 개념 수립 단계에서부터 시스템에 보안과 개인정보보호조치를 설계하는지 등을 점검하도록 한다.

• 유럽연합 <신뢰할 수 있는 인공지능 평가 목록>에서는 일반개인정보보호규정(GDPR)에 따라 의무적으로 적용되는 조치 중, 개인정보보호영향평가, 개인정보보호 책임자(DPO)의 지정, 감독 메커니즘(자격있는 직원으로 접근 제한, 데이터 접근 기록 및 수정 메커니즘 등), 개인정보 중심설계 및 기본설정, 개인정보 최소화, 동의철회권, 거부권, 잊힐 권리를 시스템 개발에 구현하였는지, 인공지능 시스템 수명주기 동안 개인정보보호 영향을 고려했는지 등에 대한 점검을 포함하고 있다.

(2) 데이터

Q2-1-3. 학습, 검증, 테스트 등 인공지능의 개발 과정에 사용되는 데이터셋에 대한 정보, 예를 들어 데이터셋의 출처, 구조와 유형, 사전 처리 과정 등에 대한 정보를 확보하고 있습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음

설명 ( )

Q2-1-4. 데이터셋의 정확성, 완전성, 최신성을 확인하였습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음

이를 검토하기 위해 사용한 방법은 무엇입니까.

설명 ( )

Q2-1-5. 데이터셋이 인공지능 시스템이 사용될 맥락에 적합하도록 인구집단별 다양성과 대표성을 갖추었는지 확인하였습니까.

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음

이를 검토하기 위해 사용한 방법은 무엇입니까.

설명 ( )

Q2-1-6. 데이터셋이 사상·신념, 건강, 인종이나 민족에 관한 정보, 생체인식정보 등 민감정보를 포함하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-7. 대리 변수를 통해 민감정보의 추정이 가능한지 여부를 검토하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-8. Q2-1-3 ~ Q2-1-7의 질의 전체 혹은 일부에 대한 확인이 불가능하거나 이를 확인하는 것이 불필요하다고 판단하는 경우, 그 이유는 무엇입니까. 또한 그러한 경우 데이터셋의 편향성을 방지할 수 있는 다른 방안은 무엇입니까.

설명 ( )

인공지능 인권 가이드라인은 “학습 데이터의 수집과 선정, 알고리즘의 설계와 활용방향 설정 등 인공지능 개발 전 과정에 걸쳐 편향이나 차별의 요소가 배제될 수 있도록 점검하는 절차를 마련”하도록 하고 있으며, 여기에는 “학습 데이터의 개별 요소를 검사하고 차별적 영향을 초래할 수 있는 데이터를 조정하는 등의 조치가 포함”되어야 한다. 특히, “학습용 데이터가 인공지능의 판단에 직접적인 영향을 미치는 상황을 고려할 때, 학습용 데이터의 수집 단계부터 차별적 요소를 통제하고 데이터 편향성을 최소화하여 인공지능을 통한 의사결정이 특정 집단에 부정적 영향을 미치지 않도록 해야” 한다.

이를 위해서는 인공지능 학습 과정에 사용되는 데이터의 투명성이 보장되어야 한다. 즉, 데이터셋의 출처가 어디인지, 데이터의 구조와 유형은 어떠한지, 사전에 어떠한 처리 과정을 거쳐 데이터셋이 구축되었는지 확인·점검이 가능해야 한다. 더불어 데이터셋이 정확성, 완전성, 최신성을 갖춰야 한다. 현실을 반영하지 못하는 오래된 데이터나 오류가 있는 데이터가 있을 경우, 인공지능의 성능에 영향을 미치고 편향을 야기할 수 있기 때

문이다. 인공지능 학습 과정에서의 편향이나 오류를 방지하기 위해서는 데이터셋 구축 단계에서 편향성이 최소화되도록 인구집단별 다양성과 대표성을 갖추었는지 확인할 필요가 있다. 물론 이는 한 사회의 인구집단을 언제나 고르게 반영해야 한다는 의미는 아니다. 다양성과 대표성은 인공지능 시스템의 활용 분야나 목적에 따라 달라질 수 있다. 예를 들어, 아동 대상 서비스를 목적으로 하는 인공지능 시스템이라면 성인 데이터보다는 아동 관련 데이터를 활용할 가능성이 크지만, 성별, 지역, 인종 등 다른 요인과 관련해서는 대표성을 갖도록 구축되어야 한다.

데이터셋에 민감정보가 포함되어 있는지 확인할 필요가 있다. 민감정보는 정보주체의 사생활을 현저히 침해할 우려가 있는 정보이므로 특별한 보호를 필요로 하기 때문이다. 국내 개인정보 보호법은 사상·신념, 노동조합·정당의 가입·탈퇴, 정치적 견해, 건강, 성생활 등에 관한 정보, 유전정보, 범죄경력자료, 생체인식정보, 인종이나 민족에 관한 정보 등을 민감정보로 규정하고 있다. 민감정보를 처리하기 위해서는 법령에서 허용한 경우이거나 별도의 동의를 받아야 한다. 인공지능 시스템에서 민감정보가 활용될 경우, 특히 취약하거나 소외된 집단에 차별적 영향이 없을지 주의할 필요가 있다. 또한 직접적으로 민감정보를 활용하지 않더라도 다른 개인정보를 통해 민감정보를 유추할 수도 있다. 예를 들어, 특정 지역에 특정 국가의 외국인이 모여 산다면 지역 정보가 인종이나 민족에 관한 정보의 대리 변수가 될 수 있는 것이다. 민감정보의 사용 및 이를 추정할 수 있는 대리변수의 사용 여부에 대해서는 명확히 파악할 필요가 있다.

물론 인터넷 상의 대량의 정보를 추출하여 데이터셋으로 활용하는 경우도 있고, 이럴 경우 데이터셋의 검증 자체가 사실상 불가능해질 수 있다. 데이터셋의 편향을 방지하기 위한 다른 방안이 있을 수도 있다. 혹은 데이터셋과 관련한 위의 질의가 해당 인공지능 시스템이 인권에 미치는 영향과 관련이 없는 경우도 있을 수 있다. 이 경우 데이터셋 검증이 불가능한 이유나 데이터셋 편향이 중요하지 않은 맥락, 혹은 편향을 방지하기 위한 다른 방안에 대해 설명한다면, 해당 인공지능 시스템의 영향을 평가하는데 도움이 될 것이다.

【참고】

• 네덜란드 <기본권 알고리즘영향평가>는 데이터의 원천 및 품질(알고리즘에 대한 입력으로 어떤 유형의 데이터가 사용되고 어떤 소스에서 가져왔는지, 데이터의 품질과 신뢰성이 충분한지), 데이터 편향성 및 가설(데이터에 어떤 가설과 편향이 반영되어 있으며, 알고리즘의 결과물에 미치는 영향이 어떻게 완화되는지, 학습데이터가 사용되는 경우 데이터가 알고리즘이 사용될 맥락을 대표하는지)을 검토하도록 한다(2A.2 및 2A.3).

• 캐나다 정부 <알고리즘영향평가> 도구는 데이터의 소스와 유형은 어떠한지(제9장), 데이터 품질 관리를 위한 문서화된 절차를 두고 해당 정보가 공개되고 있는지(제11장) 점검하도록 한다.

• 유럽연합 <신뢰할 수 있는 인공지능 평가 목록>은 데이터에서 최종사용자 및 주체의 다양성과 대표성을 고려했는지를 질의한다(요구사항 #5).

• 금융위원회 <금융분야 AI 개발·활용 안내서>는 데이터 품질 개선을 위해 학습데이터의 출처와 안정적인 데이터 수집 여부 점검, 데이터의 대표성·정합성을 체크, 데이터 최신성 확보, 학습데이터의 출처, 사전처리, 가공 등의 주요 과정 문서화를 점검하도록 한다(나-1).

(3) 알고리즘의 성능과 신뢰성

Q2-1-9. 인공지능 시스템 외의 다른 대안 혹은 채택된 알고리즘(또는 사전학습 모델 가중치) 외에 다른 대안에 대한 검토가 있었습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

인공지능 시스템에 사용된 알고리즘(또는 사전학습 모델 가중치)이 목적 달성에 적합한 이유는 무엇입니까.

설명 ( )

Q2-1-10. 인공지능 시스템이 의도한대로 작동하는지 성능을 측정하기 위한 지표와 방법을 갖고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

인공지능 시스템이 의도한대로 작동하지 않을 가능성이 크다면 이를 신뢰하기 힘들 것이다. 물론 그렇다고 해당 인공지능이 곧바로 인권 침해를 야기하는 것은 아니나, 애초에 의도된 목적에 부합하는 결과물을 생산하지 못한다면 효율성이 떨어지거나 사용자나 영향을 받는 사람들의 인권에 부정적 영향을 미치지 않도록 통제하기 힘들 수 있다. 따라서 인공지능 시스템이 실제 의도한대로 작동하는지 성능을 측정하기 위한 지표와 방법을 갖고 있어야 한다. 또한 인공지능 시스템에 사용된 알고리즘을 개발하거나 외부의 알고리즘을 채택할 때부터 해당 알고리즘 혹은 사전학습 모델 가중치가 목적 달성에 적합한지, 다른 대안적 수단은 없는지 확인할 필요가 있다.

Q2-1-11. 정확도와 오류율 등 성능의 수준은 의도한 목적에 적합한 정도로 설정되었습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

인공지능 시스템의 정확도와 오류율 등 성능은 어떻게 측정합니까.

설명 ( )

인공지능 시스템의 정확도와 오류율은 성능 측정을 위한 지표의 하나다. 위양성, 위음성의 수준을 어떻게 설정할지에 따라 인공지능 시스템의 신뢰도에 영향을 미친다. 그런데 이는 인공지능 시스템이 활용되는 목적에 따라 다르게 설정할 필요가 있다. 예를 들어, 일반적인 이용자를 위한 인공지능 번역의 경우에는 어느 정도의 오류가 수용가능하다. 감염병 감염 여부에 대한 테스트의 경우 목표가 전염병의 차단에 있다면 다소 위양성이 발생하더라도 위음성을 줄이는 방향으로 설정이 될 것이다. 따라서 인공지능의 의도된 용도 혹은 목적에 따라 적절한 정확도와 오류율의 수준을 설정하는 것이 중요하다. 물론 인공지능 시스템의 성능 지표가 정확도와 오류율만 있는 것은 아니다. 이진 분류가 아닌 회귀나 생성 과제와 관련된 인공지능의 성능 지표는 다르게 설정될 수 있다.

【참고】

• 네덜란드 <기본권 알고리즘영향평가>는 사용되는 알고리즘의 유형(비자기지도 혹은 자기지도 학습 알고리즘), 알고리즘이 선택된 이유, 선택되지 않은 다른

대안들이 덜 적절한 이슈가 무엇인지 점검하도록 한다(2B.1). 또한, 알고리즘의 정확도, 정확도를 결정하는 평가 기준, 알고리즘 목적상 정확도 수준이 적절한지, 테스트 방법, 편향 위험에 대응하기 위한 조치, 지표의 선택과 가중치의 기초가 되는 가설이 무엇인지, 알고리즘 오류율 등을 점검하도록 한다(2B.3).

• 유럽연합 <신뢰할 수 있는 인공지능 평가 목록>은 인공지능 시스템의 정확도가 낮으면 치명적이거나 적대적이거나 해로운 결과를 초래할 수 있는지, 정확성을 모니터링하고 문서화하기 위한 일련의 단계를 마련했는지, 최종사용자가 기대하는 정확도 수준이 반영되었는지 등을 질의한다(요구사항 #2).

• 금융위원회 <금융분야 AI 개발·활용 안내서>는 인공지능 기반 모형의 성능을 평가하기 위한 지표를 선정하고 있는지, 선정한 평가 지표에 따라 목표 수준의 달성 여부를 점검하고 미달하였을 경우 조치하고 있는지 점검하도록 한다(다-1).

(4) 차별금지

Q2-1-12. 인공지능 시스템이 활용 과정에서 합리적인 이유없이 인종, 종교, 장애, 나이, 학력, 직업, 출신 지역, 언어, 정치성향, 신체조건, 외모, 피부색, 병력, 성별, 성적 지향, 사회적 신분, 경제적 지위 등 개인과 집단의 특성에 따라 특정 집단에 대한 차별을 야기하거나 혹은 기존의 차별을 악화할 가능성이 있는지 검토하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-13. 인공지능 시스템 개발 과정에서 알고리즘에 의한 구조적 차별을 사전에 방지하기 위하여, 기획, 개발, 디자인, 마케팅, 경영진 등 조직 구성원의 다양성 확보, 구성원에 대한 반차별 교육, 조직 내 인공지능 윤리 정책 수립 등의 대책을 마련하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

인공지능 인권 가이드라인은 “성별, 종교, 장애, 나이, 출신 지역, 신체조건, 피부색, 성적 지향, 사회적 신분 등 개인과 집단의 특성에 따라 편향적이고 차별적인 결과가 나오지 않도록” 하고 있으며, “학습 데이터의 수집과 선정, 알고리즘의 설계와 활용방향 설정 등 인공지능 개발 전 과정에 걸쳐 편향이나 차별의 요소가 배제될 수 있도록 점검하는 절차를 마련”하도록 하고 있다.

인공지능 학습 데이터에 내재된 차별이나 편향이 인공지능 시스템의 차별과 편향을 야기하는 주요 요인 중 하나다. 이와 함께 알고리즘의 설계 과정에서도 편향이나 차별의 요소가 포함될 수 있다. 조직 구성원의 편향성(예를 들어, 특정 성의 비율이 과도하게 높거나 성소수자를 배제하는 등)과 같은 구조적 문제가 있을 수도 있고, 사회적 고정관념(예를 들어, 업무의 성별분업에 대한 전통적 인식)이 반영된 것일 수도 있다. 인공지능 시스템이 이러한 편향이나 차별적 결과를 야기하는지 테스트하는 절차가 필요하다. 또한, 개발 단계에서 구조적인 차별을 방지하기 위하여 조직 구성원의 다양성을 확보하거나 조직 구성원에 대한 교육 프로그램을 마련하는 것이 바람직하다.

참고로 국가인권위원회법은 합리적인 이유 없이 성별, 종교, 장애, 나이, 사회적 신분, 출신 지역(출생지, 등록기준지, 성년이 되기 전의 주된 거주지 등을 말한다), 출신 국가, 출신 민족, 용모 등 신체 조건, 기혼·미혼·별거·이혼·사별·재혼·사실혼 등 혼인 여부, 임신 또는 출산, 가족 형태 또는 가족 상황, 인종, 피부색, 사상 또는 정치적 의견, 형의 효력이 실효된 전과(前科), 성적(性的) 지향, 학력, 병력(病歷) 등을 이유로 한, 고용 등 일정 분야에서의 차별 행위를 “평등권 침해의 차별행위”로 규정하고 있는데, 특정 집단에 대한 차별을 판단할 때 이러한 특성 들을 고려할 수 있다.

【참고】

• 유럽연합 <신뢰할 수 있는 인공지능 평가 목록>은 인공지능 시스템이 성별,인종,피부색,민족적 또는 사회적기원, 유전적 특징, 언어, 종교 또는 신념, 정치적 또는 다른 의견, 소수민족, 재산, 출생, 장애, 나이 또는 성적지향에 기초하여 사람들을 부정적으로 차별하는지, 인공지능 시스템의 개발, 배포 및 사용단계에서 잠재적인 부정적인 차별(편향)을 테스트하고 모니터링하는 절차를 마련했는지 질의한다. 또한, 인공지능 설계자와 개발자가 인공지능 시스템을 설계하고 개발할 때 주입할 수 있는 편견을 더 잘 인식할 수 있도록 교육할 수 있는 방안을

마련했는지 질의한다(요구사항 #5).

• 영국 <NMIP 알고리즘영향평가>는 통상적인 윤리적 고려 사항의 하나로 이 프로젝트가 특정 커뮤니티에 대한 불평등 또는 불법적인 차별의 생성 또는 악화로 이어질 수 있는지 질의한다(2.a).

• 과학기술정보통신부 <2022 신뢰할 수 있는 인공지능 개발 안내서(안)>은 요구사항으로 인공지능 모델의 편향 제거(요구사항 06), 인공지능 시스템 구현 시 발생가능한 편향 제거(요구사항 10)를 규정하고 있다.

(5) 설명가능성과 투명성

Q2-1-14. 해당 인공지능 시스템이 특정한 결정(출력)을 내리는데 관련된 요소를 추적할 수 있도록 관련된 정보(예를 들어, 결정의 내역이나 시스템에 대한 모든 변경사항 등에 대한 로그기록)가 기록되고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-15. 해당 인공지능 시스템이 특정한 결정(출력)을 내린 이유나 근거에 대해 사용자 혹은 영향을 받는 이해관계자에게 설명할 수 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

인공지능 시스템의 특성을 설명하는 용어 중 하나가 '블랙박스'이다. 즉, 특정 입력으로부터 특정한 출력이 산출될 때 정확히 어떠한 이유로 그러한 출력이 산출되었는지 알 수 없다는 것이다. 설명가능한 인공지능 개발에 대한 여러 노력에도 불구하고, 전문가들은 인공지능 시스템이 어떠한 결정을 내린 이유를 설명하는 것이 쉽지 않다고 말한다.

그러나 인공지능의 자동화된 결정이 특정 개인에게 중대한 영향을 미칠 경우, 그러한 결정이 내려진 합당한 이유를 제시할 수 없다면 책임성 및 책무성의 문제가 발생할 수

있다. 특히 시민에 대한 책무성을 가지고 있는 공공기관이 인공지능 시스템을 도입할 때에 이 문제가 특히 중요해진다. 예를 들어, 인공지능에 의해 사회보장 서비스 지급 여부가 결정될 경우, 왜 어떤 사람에게는 지급되지 않았는지 설명할 수 있어야 할 것이다. 인공지능 시스템이 어떠한 논리에 따라 특정한 결과를 산출하였는지, 또는 특정한 결정을 한 이유가 무엇인지 설명할 수 있어야 한다. 만일 설명할 수 없다면, 대상자에게 설명을 해야하는 분야에서는 활용할 수 없는 것이다.

물론 모든 인공지능의 결과물이 설명가능해야 한다는 의미는 아니다. 예를 들어, 인공지능 바둑과 같이 그 이유를 정확히 몰라도 인공지능 시스템의 사용자나 다른 주체의 권리에 큰 영향을 주지 않을 수도 있다. 그러나 인권영향평가의 대상이 되는 인공지능은 인권에 부정적 영향을 미치는 고위험 인공지능일 가능성이 크며, 따라서 영향을 받는 사람에게 그 이유를 설명할 수 있어야 한다.

인공지능의 추적가능성 역시 마찬가지다. 설명가능성이 인공지능의 결과물에 영향을 받는 사람에게 그 결정의 근거에 대해 설명할 수 있어야 하는 것이라면, 추적가능성은 인공지능 시스템의 작동 과정을 기록하여 어떤 문제가 발생했을 때 사후에 문제를 야기한 요소를 추적할 수 있음을 의미한다. 물론 딥러닝과 같이 특정한 결과를 산출하기 위한 요소나 과정의 추적이 힘든 프로세스도 있을 수 있으나, 이를 둘러싼 다른 작동 과정은 기록될 수 있다. 예를 들어, 인공지능 시스템의 버전 등 모든 변경 사항, 인공지능이 내린 모든 결정(결과물)의 내역과 이와 관련된 정보 등이 로그로 기록될 수 있다. 인공지능의 추적가능성이 언제나 보장되어야 하는 것은 아닐지라도, 높은 수준의 책무성과 책임성이 요구되는 분야일수록 적절한 ‘감사(audit)’를 위해서라도 일정한 수준의 추적가능성이 보장될 필요가 있다. 어느 정도의 추적가능성이 보장되어야 하는지는 해당 인공지능 시스템이나 서비스가 활용되는 맥락에 따라 달라질 것이다.

Q2-1-16. 설명가능성 여부가 인권에 영향을 미칠 경우, 해당 인공지능 시스템의 작동이나 특정한 결정의 근거에 대해 기술 전문가가 아닌 이해관계자가 충분히 이해할 수 있는 방식으로 설명할 수 있습니까.
□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음
설명 ( )

Q2-1-17. 정확도와 오류율 등 인공지능 시스템의 성능, 어떤 결정을 내리는데 사용되는 매개변수 및 가중치, 적절한 사용법, 장점과 한계 등에 대해 사용자가 이해할 수 있는 방식으로 충분한 정보를 제공하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

인공지능 시스템의 설명가능성과 투명성은 단지 특정 결정의 이유나 근거에 대해 설명할 수 있는지만을 의미하지 않는다. 기술 전문가가 아닌 사용자나 영향을 받는 이해관계자가 이해할 수 있도록 쉬운 설명을 제공하고 있는지와 같은 소통적 측면을 포함한다.

또한, 인공지능 시스템의 투명성 보장을 위해서는 특정 결정의 이유뿐만 아니라 인공지능 시스템에 대한 전반적인 정보, 예를 들어 시스템의 성능(정확도, 오류율 등), 어떤 결정을 내리는데 사용되는 매개변수 및 가중치, 적절한 사용법, 장점과 한계 등에 대해 사용자에게 충분한 정보를 제공하는 것도 필요하다.

Q2-1-18. 인공지능 시스템의 소스코드가 공개되거나 이를 요구하는 이해관계자에게 제공될 수 있습니까. 소스코드가 제공될 수 있다면, 누구에게 어떤 조건으로 제공됩니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

특히 공공부문에서 활용하는 인공지능 시스템의 경우, 국가기밀 등 예외적인 경우가 아니라면 특정 등록소에 공개하는 것을 검토할 수 있다. 공공의 지원을 받은 결과물이라면 공공재적 성격을 갖기 때문에 모두가 그 결과물을 향유할 수 있도록 보장할 필요가 있다. 민간기업 역시 오픈소스로 공개하는 경우도 있는데, 그렇게 했을 때 여러 사람이 검증할 수 있고 혹은 이를 활용하여 또 다른 혁신이 가능할 수 있기 때문이다. OpenAI와 같이 이미 인공지능 분야에서도 모델들을 오픈소스로 공개하고 있는 사례가 많다. 기업의 영업비밀 등과 충돌할 경우, 일정한 조건 (예를 들어, 알고리즘에 대한 감사가 필요

한 경우 등) 하에 관련 전문가에게만 접근할 수 있는 방식을 적용할 수도 있다.

【참고】

• 네덜란드 <기본권 알고리즘영향평가>는 알고리즘이 어떤 데이터를 기반으로, 무엇을 어떻게 하는지 설명할 수 있는지, 어떤 집단에 대해 알고리즘이 설명될 수 있어야 하는지, 해당 집단에게 충분히 이해할 수 있는 방식으로 설명될 수 있는지 점검하도록 한다.

• 유럽연합 <신뢰할 수 있는 인공지능 평가 목록>은 투명성과 관련하여, 전체 수명주기 동안 인공지능 시스템의 추적가능성을 해결하는 조치(입력 데이터 품질에 대한 지속적인 평가, 인공지능 시스템이 결정을 내리는데 사용된 데이터의 추적 등)를 취했는지, 사용자가 인공지능의 결정을 이해하는지 여부를 지속적으로 조사하는지, 인공지능시스템에 의해 생성된 결정의 목적, 기준 및 제한 사항에 대해 사용자에게 알리는 메커니즘을 설정하였는지 등을 질의한다(요구사항 #4).

• 캐나다 「자동화된 의사결정 훈령」은 대외비 등 일정한 경우를 제외하고는 정보기술관리지침에 명시된 요건에 따라 캐나다 정부가 소유한 사용자 정의 소스 코드를 공개하도록 하고(6.2.6), 공개된 소스 코드에 대하여 적절한 접근 제한을 결정하도록 한다(6.2.7).

(6) 자동화 정도와 인간의 개입

Q2-1-19. 사람과 상호작용하는 인공지능 시스템의 경우, 인공지능 시스템의 사용자 혹은 상호작용하는 사람에게 상대방이 사람이 아니라 인공지능 시스템이라는 사실, 혹은 자신이 받은 결과물이나 결정이 인공지능 시스템에 의한 것이라는 점을 적절하게 알릴 수 있는 조치를 취하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-20. 인공지능 시스템이 영향을 받는 이해관계자가 인지할 수 없도록 은밀하게 작동할 수 있는 경우(예를 들어, 원격에서 작동하는 얼굴인식 시스템이 대상자 모르게 얼굴인식을 통해 신원을 파악하는 경우) 영향을 받는 당사자가 인지하지 못

하는 방식으로 인공지능 시스템이 작동하지 않도록 당사자에게 적절하게 알릴 수 있는 조치를 취하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

인공지능 인권 가이드라인은 ‘투명성과 설명의무’와 관련하여, “자동화된 의사결정에 의하여 영향을 받는 당사자는 그 결정의 이유에 대하여 설명을 듣고, 당사자 진술을 할 수 있으며, 이의를 제기할 수 있어야 한다”고 규정하고 있다. 또한, “완전히 자동화된 의사결정이 이루어진 경우에는 당사자가 해당 방식을 거부하거나 인적 개입을 요구할 수 있는 권리를 보장받아야” 한다고 하고 있다. 그런데 설명을 제공받고 이의를 제기할 수 있기 위해서는 우선 자신이 인공지능과 상호작용하고 있음을, 혹은 자신에게 어떠한 결정을 내린 주체가 인공지능이라는 사실을 인지할 수 있어야 한다. 이것은 인간이 인공지능으로부터 자율성과 존엄성을 보장받기 위한 최소한의 조건이다. 그렇지 않으면, 인공지능 시스템에 의해 의식이나 행동이 ‘조작’되거나 혹은 감시될 위험이 있다. 이러한 문제를 방지하기 위하여 인공지능 시스템의 사용자나 영향을 받는 사람에게 인공지능과 소통하고 있다는 사실을 적절한 방식으로 알릴 필요가 있다.

한편, 공공 공간에서 개인의 얼굴을 식별하여 수배자의 얼굴과 대조하는 인공지능 얼굴인식 시스템과 같이 어떠한 인공지능 시스템은 개인이 인지하지 못하는 사이에 개인을 식별하거나 감시할 수 있고, 또는 그러한 감시를 목적으로 도입될 수도 있다. 이러한 인공지능 감시 시스템을 도입할 수 있는 요건이 무엇인지에 대해서는 별도의 논의가 필요하지만, 이러한 시스템은 인간의 자율성과 인권을 심각하게 침해할 가능성이 크다. 설사 제한적인 조건하에서 도입될 수 있다고 하더라도, 권리 남용을 방지할 수 있는 안전조치가 취해져야 하며, 따라서 이러한 시스템의 위험성이 개발 단계에서부터 인지될 필요가 있다.

Q2-1-21. 인공지능 시스템의 결과물에 기반한 결정에서 인간의 역할과 인간이 재량권을 갖고 개입할 수 있는 범위와 절차가 정의되어 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-22. 인공지능 시스템이 의도한대로 작동하지 않을 경우, 인공지능 시스템의 운영자 혹은 사용자는 언제든지 시스템을 정지시킬 수 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

인공지능 시스템은 인간을 대신하여 완전히 자동화된 방식으로 어떠한 결정을 내릴 수도 있고, 인간 의사결정자를 보조하는 역할을 할 수도 있다. 인공지능 시스템의 자동화 정도, 그리고 인간 사용자의 개입 권한 혹은 재량권의 수준은 책무성에 영향을 미친다. 인간에게 중대한 영향을 미치는 의사결정일수록, 인간의 개입없이 인공지능 시스템에 의해 자동화된 방식으로 이루어진다면 인간의 자율성을 침해하고 책무성이 취약해질 위험이 크다. 따라서 인공지능 시스템이 활용되는 분야나 목적에 따라 적절한 수준에서 인간이 재량권을 갖고 개입할 수 있는 여지를 마련할 필요가 있다. 또한 그 영향을 받는 당사자 주체는 자신이 동의하지 않는 어떠한 의사결정에 대해서 설명을 듣고 이의를 제기할 수 있는 권리를 보장받아야 한다. (이는 제3단계에서 다룰 예정이다.) 때로 인공지능 시스템이 애초에 의도한대로 작동하지 않는다고 판단될 경우 운영자 혹은 사용자는 언제든지 시스템을 정지시킬 수 있도록 설계되어야 한다.

【참고】

• 네덜란드 <기본권 알고리즘영향평가>는 알고리즘의 결과로 어떠한 의사결정이 이루어지는지, 의사결정에서 인간은 어떠한 역할을 하는지, 알고리즘 사용의 효과는 무엇이며 낙인, 차별 등 부정적인 영향의 위험은 무엇인지, 알고리즘 의사결정의 절차와 행위자가 참여할 수 있는 방법은 무엇인지, 알고리즘이 사용되는 맥락(시기, 영역)은 어떠한지 등을 질의한다(3.1-3.5).

• 캐나다 「자동화된 의사결정 훈령」은 자동화된 의사결정 시스템이 인적 개입을 허용하도록 보장(6.3.9)할 것과 자동화된 의사결정 시스템을 생산하기 전에 적절한 수준의 승인을 획득할 것(6.3.10)을 요구한다. 이와 관련하여, 캐나다 정부의

<알고리즘영향평가> 도구는 해당 시스템이 의사결정자를 보조하는 데만 사용되는지, 인간에 의해 이루어질 수 있는 의사결정을 대체하는지, 판단이나 재량권이 필요한 인간의 의사결정을 대체하는지, 시스템을 개발한 부서가 아닌 다른 부서에서 이 시스템을 사용하는지를 검토하도록 하고 있다.

• 유럽연합 <신뢰할 수 있는 인공지능 평가 목록>은 인공지능 시스템이 인간인 최종 사용자와 상호작용하고, 안내하며, 최종사용자가 결정을 내리도록 설계되었는지, 최종사용자에게 인공지능 시스템과 상호작용하는지 여부에 대해 혼란을 일으킬 수 있는지, 최종사용자의 과도한 의존도를 발생시켜 인간의 자율성에 영향을 미칠 수 있는지, 의도하지 않고 바람직하지 않은 방식으로 최종사용자의 의사결정 과정을 방해하여 인간의 자율성에 영향을 미칠 수 있는지, 인간의 애착을 형성하거나 중독적인 행동을 자극하거나 사용자의 행동을 조작할 위험이 있는지 질의한다. 또한, 인공지능 시스템이 자율적인 시스템인지, 인간에 의해 감독되는 방식은 어떠한지, 최종사용자에 대한 역효과를 탐지하고 대응하는 메커니즘을 수립했는지, 필요할 때 작업을 안전하게 중단할 수 있는 "중지 버튼"이나 관련 절차를 보장하는지 질의한다(요구사항 #1).

• 영국의 <NMIP 알고리즘영향평가>는 프로젝트가 동의와 자율성을 어떻게 고려하는지, 감시 증가와 관련된 위험이 있는지(예를 들어, 시스템의 의도된 수혜자에게 시스템 사용에 대해 어떻게 알리는지), 시스템이 감시가 증가하는 것으로 해석될 수 있는지를 질의한다(2.b).

(7) 보안

Q2-1-23. 인공지능 시스템의 특성과 활용되는 분야 등을 고려했을 때, 인공지능 시스템 보안에 대한 가능한 위협이 무엇이고 보안이 침해되었을 경우 발생할 수 있는 결과 혹은 해악은 무엇입니까.

설명 ( )

Q2-1-24. 인공지능 시스템의 학습 및 테스트에 활용되는 데이터셋에 대해 충분한 안전조치가 적용되었습니까. 데이터 오염 등 데이터에 대한 다양한 유형의 공격에 대한 대응이 고려되었습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-25. 인공지능 시스템의 전체 수명주기 동안 발생할 수 있는 잠재적인 공격에 대비하여 무결성, 가용성, 기밀성, 견고성 등 보안에 요구되는 요소를 보장하기 위한 조치를 취했습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

인공지능 시스템이 보안 위협에 취약하면 그만큼 시스템이 애초에 의도한 바와 다르게 작동하고 인간에게 피해를 야기할 가능성이 커진다. 따라서 인공지능 시스템의 보안은 인공지능의 신뢰성을 위해서도 중요하고 인권 측면에서도 중요하게 다뤄져야 한다.

인공지능 시스템의 보안 취약점 혹은 외부의 보안 위협은 시스템의 특성이나 활용되는 분야, 목적 등에 따라 달라질 수 있다. 예를 들어, 경제적, 사회적으로 중요한 인공지능 시스템일수록 보안 위협에 노출될 가능성이 클 것이다. 인공지능 시스템의 보안 위협은 데이터셋에 대해서도 발생할 수 있고 알고리즘 및 시스템에 대해서도 발생할 수 있다. 활용되는 데이터셋의 종류에 따라서도 위협의 종류와 정도가 달라질 수 있다. 따라서 인공지능 시스템의 특성이나 활용되는 분야, 목적 등을 고려하여 보안 취약점이 될 수 있는 요소, 외부의 위협 요인 등을 파악하고 이에 따라 적절한 보안 조치를 취할 필요가 있다.

【참고】

• 유럽연합 <신뢰할 수 있는 인공지능 평가 목록>은 인공지능 시스템이 사이버 보안에 대해 인증을 받았는지, 사이버 공격에 얼마나 노출되어 있는지, 다양한 유형의 취약점과 공격을 고려했는지, 시스템의 무결성, 견고성 및 전반적인 보안을 보장하기 위한 조치를 취했는지 등을 질의한다(요구사항 #2).

• 과학기술정보통신부 <2022 신뢰할 수 있는 인공지능 개발 안내서(안)>은 요구사항으로 오픈소스 라이브러리의 보안성 및 호환성 확보(요구사항 05), 인공지능 모델 공격에 대한 방어 대책 수립(요구사항 07)을 규정하고 있다.

(8) 접근성

Q2-1-26. 인공지능 시스템이 언어, 나이, 장애, 신체적 조건 등에 상관없이 누구나 사용할 수 있도록 인공지능 시스템의 인터페이스가 보편적 설계 원칙에 따라 설계되었습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-27. 보편적 설계 원칙에 따라 설계하지 않은 합리적인 이유가 있다면 그것은 무엇입니까.

설명 ( )

인공지능 시스템 역시 정보통신 시스템의 하나이며 장애인도 접근, 활용할 수 있도록 설계되어야 한다. 그렇지 않으면 장애인, 아동이나 노인 등 사회적 취약계층에 대한 차별을 야기하고 정보격차를 심화할 수 있기 때문이다. 장애인도 보조적 수단을 이용하여 접근 사용할 수 있도록 사용자 인터페이스를 설계해야 한다. 여기에는 인공지능 시스템이 상호작용하는 주체들이 이해하기 쉬운 방식으로 표현하는 것도 포함된다.

물론 성인용 서비스를 제공하는 인공지능 시스템의 경우 연령 제한을 하거나 각 장애별 맞춤형 서비스를 위한 인공지능 등 보편적 설계 원칙을 적용하지 않은 합리적인 이유가 있을 수 있으므로 그럴 경우 그 근거를 설명하도록 한다.

【참고】

• 유럽연합 <신뢰할 수 있는 인공지능 평가 목록>은 접근성과 관련하여, 인공지능 시스템의 사용자 인터페이스가 특별한 필요나 장애가 있는 사람들이 사용할 수 있는지, 계획 및 개발단계에서 보조기술이 필요한 최종사용자와 협의했는지, 개발 과정의 모든 단계에서 보편적 설계 원칙을 고려했는지 질의한다(요구사항 #5).

(9) 라이선스

Q2-1-28. 인공지능 시스템의 전체 혹은 일부 소프트웨어를 외부에서 개발된 것을 사용할 경우, 인공지능 시스템에 의한 잠재적 인권 침해를 방지, 완화하기 위하여 필요한 경우 알고리즘 혹은 소스코드 등 소프트웨어를 적절하게 수정, 변경할 수 있는 권한에 대해 외부 개발업체와 명확한 합의가 이루어져 있습니까. ☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음 설명 ( )

인공지능 인권영향평가는 인공지능 시스템 개발자뿐만 아니라, 판매자나 사용자에 의해 수행될 수 있다. 인공지능 시스템이 외부 업체에 의해 개발되었을 경우, 인공지능 시스템의 사용자가 인권 침해 위험을 완화하기 위하여 필요할 경우 적절하게 통제할 수 있는 권한이 있는지가 문제가 된다. 특히 공공기관의 경우 민간업체의 인공지능 시스템을 조달하여 사용하게 되지만, 인공지능 시스템 적용 대상인 시민에 대한 책무성을 보장하기 위해 사용하는 인공지능 시스템에 대한 적절한 통제 권한을 가질 필요가 있다. 이를 위해서는 인공지능 시스템이나 서비스의 원 개발자 혹은 소유자로부터 해당 소프트웨어의 사용이나 변경의 범위 및 제한에 대해 명확하게 합의할 필요가 있다. 이러한 권한이 제한될 경우 인공지능 시스템을 사용하는 조직의 책무성에 어떠한 영향을 미치는지, 해당 시스템을 계속 사용하는 것이 바람직한 것인지에 대한 판단이 필요하다.

【참고】

• 네덜란드 <기본권 알고리즘영향평가>는 알고리즘이 외부에서 개발된 경우 알고리즘의 소유권 및 관리 권한에 대해 명확한 합의가 이루어졌는지, 그 합의의 내용은 무엇인지 점검하도록 한다(2B.2).

나. 인권에 미치는 영향 및 심각도

(1) 영향을 받는 인권

Q2-2-1. 인공지능 시스템이 도입, 활용될 경우 시민들의 인권에 미칠 수 있는 부정적인 영향 혹은 위험은 무엇입니까. 누구의 인권이 어떤 방식으로 침해될 수 있습니까.

설명 ( )

※ 인권이란 “대한민국의 헌법과 국제인권규범에서 인정하는 인간으로서의 존엄과 가치 자유와 권리”를 의미합니다. 예를 들어, 세계인권선언, 사회권규약, 자유권규약 등이 국제인권규범에 포함됩니다. 국제인권규범에 대한 상세한 사항은 국가인권위원회 홈페이지[130]를 참고하시기 바랍니다.

인공지능에 의해 어떠한 인권이 침해될 수 있는지 검토하기 위해 아래 표의 사례를 참고해주시기 바랍니다. 이 사례들은 극히 일부분의 예시일 뿐이며 침해될 수 있는 모든 인권을 의미하는 것은 아닙니다.

인권영향평가의 기준은 국제 및 국내 인권기준이다. 유엔 <기업과 인권 이행지침> 12문은 ‘국제적으로 인정된 인권’을 “기업의 인권 존중 책임에서 인권은 국제적으로 인정된 인권을 지칭한다. 최소한 국제권리장전과 국제노동기구의 ‘노동에 있어서의 기본원칙과 권리에 관한 ILO 선언’에서 설명하고 있는 기본적인 권리에 대한 원칙을 포함한다”고 규정하고 있다. 즉, 세계인권선언, 자유권규약, 사회권규약 및 국제노동기구의 8개 핵심 협약 등이 포함된다. 이 외에도 여성차별철폐협약, 아동권리협약, 장애인권리협약 등 특별한 보호를 필요로 하는 사람들의 인권에 대한 협약들도 포함된다. 이러한 국제인권기준을 반영한 국내 헌법도 기준이 될 수 있다.

인공지능 시스템의 활용 분야에 따라 침해 우려가 있는 인권의 종류가 달라질 수 있

130) 국가인권위원회 홈페이지, “국제인권규범”. <https://www.humanrights.go.kr/site/program/board/basicboard/list?boardtypeid=7065&menuid=001003007007 (접근일: 2022. 11. 1)>.

다. 예를 들어 범죄 감시 목적의 인공지능이라면 사생활의 권리나 집회 시위의 자유에 영향을 미칠 가능성이 크고, 소셜네트워크 플랫폼에서 사용하는 콘텐츠 관리 목적의 인공지능이라면 표현의 자유에 대한 영향이 클 것이다. (예를 들어, 선정적인 표현을 차단한다는 명분으로 성소수자의 콘텐츠를 차단할 수도 있고, 거꾸로 이용자들이 선호한다는 이유로 성폭력적인 콘텐츠를 노출할 수도 있다.) 어떠한 기본권에 영향을 미칠 것인지는 인공지능 시스템의 기능과 적용 분야, 활용되는 사회적 맥락 등에 따라 달라질 수 있다. 인권의 항목이 매우 폭넓기 때문에, 침해될 수 있는 모든 인권 항목을 체크리스트로 만들기는 힘들다. 인권영향평가를 수행하는 담당자는 해당 인공지능 시스템의 특성과 인권 기준을 고려하여 어떠한 인권에 부정적인 영향을 미칠 수 있는지 검토해야 한다. 이때 하나의 인권에만 영향을 미치는 것이 아니라 여러 인권(들)에 영향을 미칠 수 있으며, 영향의 정도와 방식이 다양할 수 있음을 고려해야 한다.

기획자 및 개발자가 다양한 인권적 영향에 대해 파악하는 것은 쉽지 않을 수 있다. 따라서 1단계에서 수행했던, 여러 이해관계자와, 특히 인공지능 시스템의 영향을 받는 이해관계자와의 소통과 협의를 수행하는 것이 이러한 인권적 영향에 대한 파악 및 분석에 도움이 될 수 있다.

인공지능 시스템의 목적과 적용 분야에 따라 침해 가능성이 있는 인권의 사례는 아래의 표를 참조할 수 있다. 이러한 시스템이 반드시 인권 침해적이라는 의미는 아니며, 어떤 인권과 관련이 있을 수 있는지 나타낼 뿐이다. 아래의 목록은 인공지능 시스템에 의한 인권침해 가능성의 극히 일부의 사례일 뿐이다.

[표 13] 인공지능 시스템에 의해 침해 가능성이 있는 인권

인공지능 시스템 예시	침해될 가능성이 있는 인권
노인과 장애인 등 대상자의 맥박, 혈당, 활동 등을 감지하고 말벗, 인지기능을 지원하는 돌봄로봇	개인정보자기결정권 침해
얼굴인식에 기반한 출입국 자동화 시스템	인종, 국가 등에 따른 차별 개인정보자기결정권 침해
지역별로 범죄 발생 확률을 예측하여 순찰 인력을 배치하는 인공지능 범죄 예측 시스템	인종, 지역 등에 따른 차별
인공지능 채용(면접) 시스템	성별, 연령, 장애, 용모, 출신지역 등에 따른 차별
공공 장소에서의 행인의 얼굴을 인식하여 용의자와 대조하는 원격 얼굴인식 시스템	이동의 자유 침해, 집회 및 결사의 자유 침해, 자의적 체포
아동이 사용하는 소셜네트워크서비스에서 선정적이고 자극적인 콘텐츠가 우선 노출되도록 하는 알고리즘	아동의 권리 침해 개인정보자기결정권 침해
소셜네트워크 플랫폼에서의 알고리즘 기반 콘텐츠 관리 시스템	표현의 자유 침해 정보접근권 침해
소셜네트워크 플랫폼에서 개인의 정치 성향에 기반한 정치광고 노출 시스템	자유로운 정치참여 제한 선거권 침해
고등학교의 기존 성적에 기반한 인공지능 대학입학 시스템	지역에 따른 차별 교육권 침해
사업장 내에 설치된 생체인식, 위치추적 시스템	노동자 개인정보자기결정권 및 노동3권 침해
인공지능 판결 지원 시스템	공정한 재판을 받을 권리 침해
인공지능을 통한 사회보장급여 부정수급 탐지시스템	사회보장수급권, 장애인권리 침해, 인종 및 장애 등에 따른 차별

Q2-2-2. 인공지능 시스템이 오류로 인하여 의도하지 않은 방식으로 작동할 경우 나타날 수 있는 부정적인 결과에 대해 검토한 바 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

의도하지 않은 방식으로 작동할 경우 침해되는 인권은 무엇입니까

설명 ( )

Q2-2-3. 인공지능 시스템이 의도적으로 악용될 가능성이 있습니까. 어떠한 방식으로 오용될 수 있는지에 대해 검토하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

악용될 경우 나타날 수 있는 부정적인 결과, 혹은 침해되는 인권은 무엇입니까

설명 ( )

다른 전자기술과 마찬가지로 인공지능 시스템 역시 오류에 의해 의도하지 않은 방식으로 작동하거나 혹은 의도적으로 악용될 수 있다. 인공지능 시스템이 의도된 대로 작동할 경우에도 인권에 부정적인 영향을 미칠 수 있지만, 의도된 목적과 달리 작동할 경우 역시 고려할 필요가 있다. 인권영향평가 과정에서 인공지능 시스템이 의도된 목적과 달리 작동할 경우 발생할 수 있는 부정적 영향이 무엇인지, 인권에 미치는 영향은 어떠한지 파악한다.

【참고】

• 유럽연합 <신뢰할 수 있는 인공지능 평가 목록>은 기본권과 관련하여 차별금지, 아동의 권리, 개인정보 보호, 표현과 정보의 자유, 집회결사의 자유를 존중하는지 질의한다.

• 덴마크 <디지털활동 인권영향평가>는 국제인권기준을 인권영향평가의 기준으로 제시한다. 영향 분석, 영향 심각도 평가 및 완화 조치 설계는 국제 인권 기준 및 원칙에 따라 수행되어야 한다.

• 영국 <NMIP 알고리즘영향평가>는 해당 시스템이 보급된 후 이 시스템의 사용으로

인해 발생할 수 있는 최상의 시나리오와 최악의 시나리오, 성공을 위해 필요한 사회 환경 조건 및 극복해야 할 예상되는 장애물이나 과제는 무엇인지 설명하도록 한다. 이때 이 시스템이 설계/의도된 대로 작동할때 뿐만 아니라 실패, 오류, 실수 또는 예기치 않은 동작을 처리하는 방법에 대해서도 고려해야 한다. 이 프로세스에서 식별되어야 하는 영향의 최소 개수는 없다(3. 영향 식별 및 시나리오).

• 금융위원회 <금융분야 AI 개발·활용 안내서>는 인공지능 시스템의 기획 및 설계 단계에서 고객에게 미치는 영향, 위험수준, 잠재적 피해 가능성 등이 고려되었는지 검토하도록 한다(가-1).

(2) 인권에 미치는 영향의 심각도

Q2-2-4. 인공지능 시스템이 인권에 미치는 부정적 영향의 범위가 어떠합니까. 전체 인구 혹은 어떠한 특정 집단에 대하여 어느 정도의 범위(대, 중, 소)로 영향을 미칠 수 있습니까. (부정적 영향을 받을 수 있는 인권이 여러 개인 경우 각각에 대해서 평가가 필요함. 아래 질의에 대해서도 동일함)

설명 ( )

Q2-2-5. 인공지능 시스템이 인간의 생명, 건강, 안전, 인권, 기본적 삶 등에 미치는 부정적 영향의 규모 혹은 크기가 어떠합니까. (대, 중, 소)

설명 ( )

Q2-2-6. 인공지능 시스템이 인권에 미치는 부정적 영향이 사후에 구제나 회복이 어느 정도 가능합니까. (완전히 회복 가능, 부분적으로 회복 가능, 회복 불가능)

설명 ( )

평가자들은 부정적 영향의 각 측면에서 어느 정도로 심각한지에 대해 인권영향평가 과정에서 판단할 필요가 있다. 다만, 본 인권영향평가(안)에서는 범위, 규모, 회복 불가능성의 수준을 정량적으로 구분하지는 않았다. 따라서 부정적 영향을 범위 등이 얼마나 큰지 여부에 대해 판단할 때 당연히 주관성이 개입될 수밖에 없다. 다만, 그것은 평가팀이

일방적으로 판단하는 것이 아니라 다양한 이해관계자와의 의견수렴과 협의를 통해 판단해야 한다.

피해의 범위와 관련하여 인공지능 시스템의 목적과 활용 분야에 따라 전체 인구에 어느 정도의 영향을 미치는지가 중요할 수도 있지만, 또는 특정 집단에 미치는 범위가 중요할 수도 있다. 예를 들어, 전체 인구에 대한 영향을 받는 인구의 비율은 작을 수 있지만, 외국인 노동자 등 특정 집단에 상당한 비율로 부정적 영향을 미친다면 이는 그 영향이 광범위하다고 얘기할 수 있다. 또한, 부정적 영향을 받는 대상이 취약하거나 소외된 집단이라면 부정적 영향의 심각성을 보다 높게 평가할 수 있을 것이다. 부정적 영향이 회복 가능한지 여부도 중요하다. 피해에 대한 구제나 원상 회복이 가능하지 않을 경우, 해당 인공지능 시스템의 도입에 더욱 신중할 필요가 있고, 도입을 하더라도 안전조치 마련에 보다 신경을 쓸 필요가 있다.

부정적 영향의 범위, 규모, 회복 불가능성의 수준이 모두 높게 나타난다면, 이를 예방하거나 완화할 수 있는 적절한 보완책이 마련되어야하고, 적절한 보완책이 없다면 인공지능 시스템의 도입을 철회하는 것까지 고려해야 한다.

Q2-2-7. 인공지능 시스템이 인권에 미치는 부정적 영향이 여러 개일 경우 서로 상충하는 인권이 있습니까. 또한 그 심각성 때문에 우선적인 대응이 필요한 인권은 무엇입니까.
설명 ( )

인공지능 시스템은 여러 인권에 동시에 부정적인 영향을 미칠 수 있고 영향을 미치는 방식이나 심각성은 모두 다를 수 있다. 각 인권에 미치는 심각성의 정도 역시 판단할 수 있다. 여러 권리의 상충 관계(예를 들어 표현의 자유와 프라이버시권 사이의 상충관계)가 있을 수 있기 때문이다.

【참고】

• 덴마크 <디지털활동 인권영향평가>는 인권영향평가를 할 때 인권에 미치는 범위, 규모, 회복 불가능성을 고려하도록 하고 있다. 범위는 영향이 미치는 범위가 전체 인구의 몇 % 이상인지 또는 그 영향이 확인된 집단의 몇 % 이상인지에 따라 3단계로 구분한다. 규모는 안전, 건강, 기본적인 삶 등에 미치는 부정적 영향의 심각성에 따라 3단계로 구분한다. 회복 불가능성은 인권에 미치는 영향이 사후에 구제가 가능한지, 원상 회복이 가능한지 여부에 따라 3단계로 구분한다.

• 영국의 <NMIP 알고리즘영향평가>는 해당 시스템의 잠재적 영향에 대한 모든 시나리오에서 다양한 이해관계자에의 잠재적 피해가 무엇인지, 누가 어떻게 피해를 입을 가능성이 가장 높은지 설명하도록 한다. 또한 각각의 피해에 대해서 중요도(얼마나 심각한 것인지), 긴급성, 피해 완화의 어려움, 피해의 탐지가능성을 고려하도록 한다(4. 잠재적 피해 분석).

【3단계 : 개선 및 구제】

인공지능 인권 가이드라인은 “인권에 미치는 부정적인 영향이나 편향성 및 위험성이 드러난 경우 이를 방지하거나 완화하기 위한 조치사항을 수립하여 적용” 하도록 하고 있으며 방지하거나 완화하는 조치를 취하기 전에는 그 개발과 활용을 중단하도록 하고 있다. 2단계(영향 분석 및 평가)에서 인권에 미치는 부정적 영향이 파악될 경우, 위험성을 방지하거나 완화하는 조치를 취해야 한다. 가장 바람직한 것은 파악된 위험성을 통제하기 위한 조치를 취하여 부정적 영향이 나타나지 않도록 방지하는 것이다. 완전히 방지하기 어려울 경우 부정적 영향을 최소화하기 위한 완화 조치를 취해야 한다. 그럼에도 불구하고 위험성이 잔존하여 특정한 피해를 야기했을 경우 이에 대해 이용자(혹은 시민이나 소비자)가 문제를 제기하고 가능한 침해된 권리를 복구하며, 때로는 손해를 배상받을 수 있는 절차를 마련할 필요가 있다. 즉, 인공지능 시스템으로 인한 인권 침해나 부정적 영향을 방지, 완화, 구제할 수 있는 절차를 마련해야 한다. 인권영향평가를 수행하는 담당자는 이러한 조치의 존재를 확인하고 적절성 여부를 판단하며, 이에 대해 관련 이해관계자와 협의해야 한다.

가. 방지

Q3-1-1. 데이터에 대한 개선, 알고리즘의 수정, 시스템 설계 변경 등 2단계(영향 분석 및 평가)에서 파악된 중대한 인권 위험을 방지하기 위해 어떠한 조치를 취했습니까.
설명 ( )

인공지능 시스템의 인권 침해 위험을 방지하기 위한 조치는 문제가 식별된 데이터에 대한 개선, 알고리즘에 대한 수정, 인공지능 시스템 설계의 변경 등을 포함한다. 또한 2단계 체크리스트에서 제기된 여러 조치들, 예를 들어, 개인정보 영향평가, 데이터셋의 정확성, 완전성, 최신성의 확인, 알고리즘 시스템 성능에 대한 사전 테스트, 알고리즘 공정성을 위한 구조적 조치, 설명가능성을 위한 조치, 인간의 개입 가능성 조치, 장애인 접근

성 조치 등이 부정적 영향을 방지, 혹은 완화하기 위한 조치가 될 수 있다.

【참고】

• 덴마크 <디지털활동 인권영향평가>는 영향 완화조치는 모든 인권영향을 다루도록 한다. 영향을 다루기 위한 조치의 우선 순위를 정해야 하는 경우, 인권영향의 심각도가 핵심 기준이다. 식별된 영향을 해결할 때는 “방지-감소-회복-시정”의 완화 계층 구조를 따른다.

나. 완화

Q3-2-1. 2단계(분석 및 평가)에서 파악된 인권 위험을 완전히 방지하기 곤란한 경우, 위험을 완화하기 위해 어떠한 조치를 취했습니까.

설명 ( )

Q3-2-2. 인공지능 시스템의 잔존하는 위험성에 대해 사용자 및 영향을 받는 이해관계자에게 충분한 정보를 제공하고 올바른 작동 방법에 대해 적절한 교육을 제공하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q3-2-3. 인공지능 시스템의 인권 침해 위험이 클 수 있는 특정한 사용을 허용하지 않도록 이용약관이나 여타의 집행체계에서 금지하는 절차를 취하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q3-2-4. 2단계(분석 및 평가)에서 파악된 중대한 인권 위험이 완화되지 않고 남아있을 경우 그 이유를 문서화하고 있습니까

□ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음 설명 ( )
Q3-2-5. 중대한 위험에 대한 방지 및 완화 조치를 취하기 힘들거나, 이러한 조치를 취해도 여전히 중대한 위험이 남아있을 경우 인공지능 시스템의 개발 및 활용을 중단할 계획을 갖고 있습니까 □ 예 □ 보완 필요 □ 아니오 □ 정보 없음 □ 해당 없음 설명 ( )

2단계 체크리스트에서 제기된 여러 조치들은 인권에 대한 부정적 영향을 완전히 방지하지는 못하더라도 완화할 수 있는 조치가 될 수 있다. 이에 더하여 위 체크리스트의 조치들을 검토할 필요가 있다.

여전히 남아있는 위험성에 대해 사용자에게 투명하게 설명하고 안전하게 사용할 수 있는 방법을 적절한 방식으로 교육함으로써 사용 과정에서의 위험을 최소화할 수 있다. 때로는 특정한 방식으로 인공지능 시스템을 사용하지 않도록 이용약관 등을 통해 금지하는 것도 하나의 방법이다. 예를 들어, 구글의 유명인 얼굴인식 API에 대한 인권영향평가 사례에서, 구글로 하여금 이 API를 적절한 자격을 갖춘 업체(고객 화이트리스트)에게만 팔도록 제한하고, 오로지 전문적인 영화에만 사용하도록 허용되는 사용 사례를 제한하도록 권고한 것이 이에 해당한다.

중대한 위험이 남아있음에도 불구하고 추가적인 조치를 취하지 않거나 취할 수 없는 경우 그 이유를 문서화하여 남겨두어야 한다. 그래야 중대한 위험이 있는 인공지능 시스템을 사용해야 할 다른 가치가 있는지, 사용한다면 어떠한 조건에서 사용할 수 있는지 등에 대한 판단에 도움이 될 수 있다. 중대한 위험을 제거할 수 없을경우 해당 인공지능 시스템의 개발 및 활용을 중단하는 것도 선택지로 고려할 필요가 있다.

【참고】

• 유럽연합 <인공지능법(안)>은 고위험 인공지능 시스템과 관련하여 위험관리 시스템이 수립, 이행, 문서화, 유지될 것을 요구한다. 위험관리 시스템은 고위험 인공지능 시스템의 전체 생명주기를 통해 계속적이고 반복적인 과정으로 이루어진다. 적절한 위험관리 조치를 파악할 때, (a) 가능한 적절한 설계 및 개발을 통한 위험의 제거 혹은 감소, (b) 제거될 수 없는 위험과 관련하여 적절한 완화 및 통제 조치의 이행, (c) 위험에 대한 적절한 고지 및 이용자에 대한 교육 제공 등이 보장되어야 한다(제9조).

• 미 의회 <2022년 알고리즘 책무성법(안)>에서는 영향평가 요구사항의 하나로, 모든 관련 피고용인에게 유사한 자동화된 의사결정 시스템이 소비자에게 미치는 중대한 부정적 영향에 대한 내용, 업계 모범 사례 및 제안, 해당 시스템에 대한 영향평가 개발이나 수행 방식을 개선하는 것과 관련한 내용에 대해 지속적인 교육훈련을 수행할 것(sec.4.(a)(5)), (이용약관 등을 통한 적용의 금지 및 제한을 포함하여) 자동화된 의사결정 시스템의 특정한 사용 및 적용에 대한 보호막이나 한정의 필요성과 개발 가능성을 평가할 것(sec.4.(a)(6)), 자동화된 의사결정 시스템이소비자에게 미치는 중대한 부정적 영향 가능성을 식별하고 적용가능한 완화 전략을 평가할 것. 여기에는 중대한 부정적 영향 가능성에 대한 식별 및 측정, 시스템을 철수하거나 개발을 종료하는 등의 조치를 포함하여 식별된 중대한 부정적 영향 가능성을 제거하거나 합리적으로 완화하기 위한 조치, 해당 영향이 완화되지 않은 상태로 남아 있는 것과 조치를 취하지 않은 이유 문서화 등이 포함된다(sec.4.(a)(9)).

• 네덜란드 <기본권 알고리즘영향평가>는 “기본권에 대한 알고리즘의 영향을 완화할 수 있는 규제 및 관리 도구”로서 알고리즘의 특정 사용에 대해 허용되지 않는 것으로 간주하거나 잠정중단(모라토리엄) 채택, 알고리즘을 다루는 행위자에 대한 행동 강령, 전문적 표준 또는 윤리 강령 도입, 데이터 윤리 인식에 대한 교육 또는 훈련 실시, 알고리즘 결과물을 기반으로 한 의사결정에서 체크리스트를 사용, 영향을 받거나 참여적인 시민에 참여 기회 제공, 알고리즘이 더 이상 바람직하지 않은 것으로 판명된 경우 알고리즘 사용을 중단하기 위한 출구 전략 개발 등을 제시한다.

다. 구제

Q3-3-1. 인공지능 시스템의 결정에 의해 영향을 받는 사람이 인공지능 시스템의 결정에 이의를 제기하거나 침해된 권리의 구제를 요구할 수 있는 절차를 마련하고, 이에 대한 정보를 누구나 쉽게 접근할 수 있도록 일반에 공개하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q3-3-2. 인공지능 시스템에 의해 영향을 받는 사람들에게 인공지능 시스템의 사용을 거부할 수 있는 선택권(옵트아웃 권리)을 제공하거나 이의를 제기할 수 있는 수단이 마련되어 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q3-3-3. 인공지능 시스템이 내린 결정에 의해 영향을 받는 이해관계자가 인공지능 시스템의 적용을 거부할 경우, 인간의 지원 혹은 인공지능이 아닌 시스템의 적용을 대안으로 제시할 것을 고려하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q3-3-4. 인공지능 시스템의 결정에 대한 이의제기나 권리구제 요구가 정당할 경우, 문제의 의사결정을 번복하거나 권리를 복구하거나 손해배상을 할 수 있는 절차가 마련되어 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

인공지능 시스템이 인권에 미칠 부정적 영향을 완전히 배제할 수 없는 이상, 부정적 영향을 받는 당사자가 인공지능 시스템의 결정에 이의를 제기하거나 침해된 권리의 구제를 요구할 수 있는 절차를 제공할 필요가 있다. 이용자들은 인공지능 시스템의 결정에 이의를 제기하고, 추가적인 정보를 제공하여 다시 결정을 내릴 것을 요구할 수 있다. 혹은 인공지능 시스템의 사용을 거부하거나, 인공지능 시스템에 의한 의사결정의 대상이 되는 것을 거부하고 의사결정에 있어서 사람의 개입을 요청할 수도 있다. 인공지능 인권

가이드라인은 “특히, 완전히 자동화된 의사결정으로만 개인에게 법적 효력 또는 생명·신체·정신·재산에 중대한 영향을 미치는 일은 제한되어야 하고, 이러한 의사결정이 이루어진 경우에는 당사자가 해당 방식을 거부하거나 인적 개입을 요구할 수 있는 권리를 보장받아야” 한다고 규정하고 있다. 이를 위해서는 앞서 본 바와 같이, 인공지능 시스템이 어떠한 결정을 한 근거에 대해 당사자에게 충분한 설명이 제공되어야 한다. 당사자의 이의제기가 정당하다고 판단이 될 경우, 해당 결정을 취소하고 침해된 권리가 있을 경우 이를 회복할 수 있는 조치를 취하며, 필요할 경우 손해배상을 받을 수 있는 권리를 보장해야 한다.

【참고】

• 영국 <NMIP 알고리즘영향평가>는 해당 시스템을 사용하거나 영향을 받는 개인이 결과에 대해 어떻게 이의를 제기할 수 있는지, 이 시스템 사용을 거부할 수 있는 옵션이 있는지 질의한다(2.e).

• 미 의회 <2022년 알고리즘 책무성법(안)>에서는 영향평가 요구사항의 하나로, 소비자에게 해당 시스템이 사용될 것이라는 점과 이로부터 제외(옵트아웃)될 수 있는 방법을 제공하는 정도를 평가하도록 한다. 또한, 해당 시스템의 투명성과 설명가능성을 평가하고, 소비자가 결정에 대해 이의제기·정정·재심을 청구하거나 해당 시스템 또는 프로세스에서 제외될 수 있는 정도를 평가하도록 한다(sec.4.(a)(8)).

• 금융위원회 <금융분야 AI 개발·활용 안내서>는 금융회사와 수탁기관으로 하여금, 소비자가 피해 발생 등에 대해 이의제기를 할 수 있는 장치 마련, 주기적인 모니터링, 소비자 피해 구제방안 마련, 손해배상 처리 절차 규정 등 소비자 피해방지 조치를 마련할 것을 규정하고 있다.

라. 이해관계자와의 의견수렴 및 협의

Q3-4-1. 인공지능 시스템의 인권 위험을 방지, 완화하고 인권을 침해받은 사람의 권리를 구제하기 위한 조치에 대해서 관련 이해관계자(1단계에서 파악한 이해관계자)의 의견을 수렴하거나 협의를 진행하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q3-4-2. 공공기관이 도입하는 인공지능 시스템의 경우, 가능한 모든 이해관계자가 참여할 수 있도록 의견 수렴을 위한 공청회를 실시하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

이해관계자와의 의견수렴 및 협의는 인권영향평가의 모든 단계에서 이루어질 필요가 있다(2단계의 경우 1단계에서의 협의를 바탕으로 평가를 하는 것이기 때문에 따로 질의를 포함하지는 않았다). 위험을 방지, 완화하고 피해를 구제할 수 있는 방안에 대해서도 해당 정책의 적절성에 대해 이해관계자, 특히 인공지능 시스템의 영향을 받을 이해관계자의 의견을 수렴하고 협의하는 것이 바람직하다. 이 단계에서도 기술에 전문적이지 않은 이해관계자들이 이해할 수 있도록 쉽게 정보를 제공할 필요가 있다. 특히 공공기관이 도입하는 인공지능 시스템의 경우, 시민 일반에 영향을 미칠 수 있으므로 공청회를 실시하는 것도 바람직하다.

【참고】

• 덴마크 <디지털활동 인권영향평가>는 영향을 받았거나 잠재적으로 영향을 받을 권리주체의 유의미한 참여가 영향평가 절차의 모든 단계(데이터 수집 및 맥락 분석; 영향 분석; 영향 예방, 완화 및 개선; 보고 및 평가 등)에 반영될 것을 인권영향평가의 핵심 요소의 하나로 제시한다.

【4단계 : 공개 및 점검】

인권영향평가 수행 담당자는 인공지능 시스템이 인권에 미치는 영향에 대한 분석 및 평가를 하고, 인권 위험을 방지, 완화하고 피해자를 구제할 수 있는 정책을 수립한 후, 이를 권장 사항으로 정리하여 최종 보고서를 작성하게 된다. 최종보고서에는 본 도구를 통해 점검한 항목과 세부적인 정보, 평가자의 평가 내용과 위험성을 완화하기 위한 기술적, 정책적 개선내용이 권장 사항으로 포함된다. 최종 보고서는 해당 조직에 보고되고 의사결정단위의 결정을 거쳐 이행이 될 것이다. 인권영향평가 보고서의 권장 사항은 인공지능 시스템의 개발부서에 전달되어 수정 및 개선이 이루어질 수 있다. 인공지능 시스템이 현장에 도입된 이후에는 인권에 미치는 영향이 실제로 어떻게 나타나는지 모니터링 해야 하며, 사전에 이를 위한 시스템 역시 마련될 필요가 있다. 인권영향평가의 보고서도 적절한 수준에서 공개하는 것이 투명성과 책무성 확보 차원에서 중요하다. 인권영향평가 과정에서의 장애물이나 문제점은 무엇이었는지 검토하여 조직의 정책에 반영한다면, 향후 다시 인권영향평가를 수행할 때 도움이 될 수 있을 것이다.

4단계의 경우 인권영향평가를 완료한 이후의 절차에 대한 것이지만, 인권영향평가 수행 담당자가 4단계를 위한 절차를 조직에서 마련하고 있는지 확인하고 마련되어 있지 않다면 필요한 정책을 수립하도록 조직에 권장할 수 있을 것이다.

가. 인공지능 시스템의 주요 요소의 공개

Q4-1-1. 인공지능 시스템이 사용하는 데이터와 알고리즘 등의 주요 요소를 일반에 공개하고 이해할 수 있는 방식으로 쉽게 설명하고 있습니까. ☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음 설명 ( )

인공지능 인권 가이드라인은 “공공기관이 개발하고 활용하는 모든 인공지능과 민간이 개발하고 활용하는 인공지능 중 개인의 생명이나 안전 등 기본적 인권에 중대한 영향을 미치는 인공지능은 사용된 데이터와 인공지능 알고리즘의 주요 요소를 일반에게 공

개하고 설명” 하도록 하고 있다. 앞서 2단계에서, 인공지능 시스템이 특정한 결정(출력)을 내린 이유나 근거에 대해 사용자 혹은 영향을 받는 이해관계자별로 설명가능성을 요구했지만, 인공지능 인권 가이드라인은 공공기관이 활용하는 인공지능 시스템 혹은 기본적 인권에 중대한 영향을 미치는 민간의 인공지능의 경우 데이터와 인공지능 알고리즘의 주요 요소를 일반적으로 공개할 것을 요구한다. 물론 이에 해당하지 않는 인공지능 시스템의 경우에도 주요 요소를 일반에 공개한다면 신뢰성을 높이는데 도움이 될 것이다.

여기서 주요 요소란 인공지능 시스템이 어떤 요소들을 기준으로 사용하여 의사결정을 하는지, 해당 결정에 있어서 각 요소들의 중요도는 어떠한지 등을 의미한다. 예를 들어, 배달 라이더에게 업무를 할당하는 인공지능 시스템이라면, 고객과의 거리, 라이더의 평점, 수락률 등 어떠한 요소가 업무 배분에 어떻게 작용하는지 라이더에게 공개할 필요가 있다는 것이다.

【참고】

• 유럽연합 GDPR에서는 프로파일링을 포함한 자동화된 의사결정이 이루어질 경우 정보주체에게 그러한 처리의 중요성 및 예상 결과와 함께, “관련 로직”에 대한 의미있는 정보를 제공할 것을 요구하고 있다.

• 유럽연합 <디지털서비스법>의 경우 온라인플랫폼으로 하여금 광고 대상 이용자를 결정하는데 사용된 주요 매개변수(가장 중요한 결정 기준과 중요성 포함)와 그 변경방법, 콘텐츠 추천에 사용되는 주요 매개변수와 그 변경방법을 공개하도록 하고 있다. 대규모온라인플랫폼의 경우, 광고가 특정 이용자 집단을 맞춤하였는지 여부 및 (특정 이용자를 배제하기 위해) 사용된 주로 매개변수를 저장소에 공개하도록 하고 있다.

• 호주 <뉴스미디어협상법>은 플랫폼이 뉴스 노출 알고리즘을 변경할 때 그 해당 사항을 언론사에 고지하도록 한다.

• 스페인 <디지털 플랫폼 유통에 종사하는 개인의 고용 상태에 관한 법률(일명 ‘라이더법’)>은 플랫폼 노동자의 근로조건, 고용과 해고 결정에 영향을 미칠 수 있는 업무 배치와 평가 관련 알고리즘과 인공지능에 관한 정보를 근로자 대표에게 공개하도록 의무화하고 있다.

나. 인권영향평가 결과 공개

Q4-2-1. 인권영향평가 보고서 전체 혹은 주요 내용을 일반에 공개합니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q4-2-2. 인권영향평가 보고서를 감독기구인 국가인권위원회에 제공하고, 효과와 한계에 대해 협의하는 절차가 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

인권영향평가가 인공지능 시스템을 개발 혹은 사용하는 조직 외부의 기관에 의해 이루어지더라도 위탁하는 조직과의 관계 때문에 독립적으로 이루어지지 못할 수도 있다. 자칫하면 피상적으로 평가가 이루어지거나 인권영향평가를 수행했다는 면책수단으로 활용될 위험이 있다. 인권영향평가의 결과를 공개하는 것은 이해관계자 및 관심있는 사람들에게 인공지능 시스템과 인권영향평가의 결과에 대한 정보를 제공하는 의미와 함께, 외부적인 평가에 노출됨으로써 인권영향평가가 제대로 이루어졌는지 검증하는 효과가 있다. 나아가 현재 의무인 것은 아니지만, 인권영향평가 보고서를 감독기구인 국가인권위원회에 제출하고, 필요할 경우 국가인권위원회가 이를 평가하고 의견을 제시하거나 완화 정책에 대해 협의할 수 있는 절차가 있다면 인권영향평가의 실효성을 높이는데 기여할 수 있을 것이다.

【참고】

• 캐나다 「자동화된 의사결정 훈령」은 알고리즘영향평가의 최종 결과를 <정부 개방 지침>에 부합하도록 캐나다 정부 웹사이트 및 캐나다 재정위원회가 지정한 기타 서비스를 통해 일반 접근이 가능한 형식으로 공개하도록 한다(6.1.4).

• 영국 <NMIP 알고리즘영향평가>는 영향평가의 결과물을 국가보건서비스(NHS) 중앙

저장소에 공개하면서, 이 영향평가에 대해 문의할 수 있도록 신청자의 연락처도 함께 공개하도록 한다. 심사를 통과한 성공적인 영향평가만 공개되지만, 그렇지 않은 신청자 역시 결과물을 공개하여, 자신이 배운 내용을 공유할 수 있도록 한다.

• 페이스북은 BSR이라는 기관에 미얀마에서의 페이스북 서비스에 대한 인권영향평가를 수행하도록 위탁했으며, 2018년 10월 인권영향평가 보고서가 공개되었다.

다. 인공지능 시스템에 대한 모니터링

Q4-3-1. 인공지능 시스템이 도입되거나 운영이 시작된 후에 그 성능과 인권에 미치는 부정적 영향, 완화 조치 및 구제 정책의 효과성을 확인하기 위해, 인공지능 시스템의 수행을 모니터링하고 기록에 남길 수 있는 메커니즘을 수립하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q4-3-2. 인공지능 시스템이 의도한 대로 작동하지 않거나 인권에 미치는 부정적인 영향이 확인되었을 때, 관련된 책임을 명확히 하고 인공지능 시스템을 개선하며 부정적 영향을 완화하기 위해 필요한 절차를 수립하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

인공지능 시스템의 실제 활용 과정에 대한 모니터링이 필요하며 이를 어떻게 수행할 것인지에 대한 메커니즘을 수립할 필요가 있다. 예를 들어, 인공지능이 목표로 하는 성능을 달성하고 있는지, 그러한 성능을 안정적으로 유지하고 있는지, 인권 위험을 방지하거나 완화하기 위해 취했던 조치들이 실제로 효과가 있는지, 애초에 예상하지 못한 인권 위험이 발생하지는 않는지 등을 모니터링해야 한다. 해당 인공지능 시스템이 야기하는 문제에 대한 외부 이해관계자(예를 들어 영향을 받는 당사자, 인권단체, 노동조합 등)의 신고를 받는 것도 하나의 방법이 될 수 있다. 이때 인공지능 시스템의 문제점을 신고하

는 이들이 취약한 지위에 있을 경우, 적절한 보호를 받을 수 있는 방안까지 고려해야 한다.

모니터링을 통해 성능의 결함이나 예상하지 못한 인권 위험 등이 발생했을 경우, 그 문제의 원인이 무엇인지, 이에 대해 누가 책임을 져야 하는지, 문제를 어떻게 개선할 수 있는지 등을 판단하고 집행하기 위한 절차가 마련되어야 한다.

【참고】

• 유럽연합 <인공지능법(안)>은 제61조에서 출시 후 모니터링의 수행을 규정하고 있다. 인공지능 시스템의 제공자는 인공지능 기술의 성격과 위험성에 비례하여 모니터링 시스템을 구축하고 문서화해야 하는데, 이는 수명주기 전반에 걸친 수행에 관한 데이터를 적극적·체계적으로 수집, 기록, 분석하고, 제공자가 고위험 시스템에 대한 요구조건을 지속적으로 준수하는지 평가할 수 있도록 한다.

• 캐나다 「자동화된 의사결정 훈령」은 자동화된 의사결정 시스템을 의도하지 않은 결과로부터 보호하고 본 지침뿐만 아니라 기관 및 프로그램 관련 법률의 준수를 확인하기 위하여 그 결과를 정기적으로 모니터링하는 절차를 개발하도록 한다(6.3.2).

• 금융위원회 <금융분야 AI 개발·활용 안내서>는 인공지능 시스템 성능이 안정적으로 유지되는지 확인할 수 있는 모니터링 절차를 마련하였는지 검토하도록 한다(라-2).

라. 인권영향평가에 대한 점검

Q4-4-1. 인권영향평가 수행의 효과와 한계를 점검하고, 개선할 수 있는 절차를 마련하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

환경영향평가나 개인정보 영향평가 등 어느 정도 제도로 정착된 영향평가와 달리, 인공지능 인권영향평가는 물론이고 인권영향평가 자체가 아직 초기 단계라고 할 수 있다.

따라서 관련된 전문가집단과 누적된 경험이 많지 않다. 인권영향평가 자체도 어느 정도는 시행착오를 통해서 개선해나갈 필요가 있다. 인공지능 인권영향평가를 조직 내에 정책화했다고 하더라도 마찬가지다. 인권영향평가 자체에 대한 점검을 통해 영향평가의 주체, 시기, 방식, 절차 등에 대해서 개선해나갈 필요가 있다. 예를 들어, 인공지능 시스템 개발의 어떤 단계에서 수행하는 것이 가장 효과가 있는지, 효과적인 인권영향평가를 위해 참여할 필요가 있는 이해관계자의 범위, 영향평가가 실제로 인권 침해의 가능성을 사전에 방지하거나 완화하는데 효과가 있었는지 등이 점검될 수 있다.

【참고】

• 미 의회 <2022년 알고리즘 책무성법(안)>에서는 해당 시스템에 대한 영향평가를 개선하는데 필요한 기능, 도구, 표준, 데이터셋, 보안 프로토콜, 이해관계자 참여 개선 및 기타 자원을 파악하도록 규정하고 있다(sec.4.(a)(11)). 또한, 시도되었으나 실행이 불가능하여 준수할 수 없었던 영향평가 요구사항 및 그 근거에 대해 문서화하도록 하고 있다(sec.4.(a)(12)).

마. 인권영향평가의 재수행

Q4-5-1. 인공지능 시스템에 대해 정기적으로 (예를 들어 1년) 인권영향평가를 수행하는 절차를 마련하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q4-5-2. 인공지능 시스템의 핵심적인 기능이 변경되거나 환경적 요인 혹은 적용 범위가 변경되었을 경우, 인권영향평가를 다시 수행하는 절차를 마련하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

인권영향평가는 특정 제품이나 서비스의 출시 전에 수행하는 절차이기도 하지만, 일반적인 인권영향평가를 비롯하여 인공지능 인공영향평가 관련 대부분의 해외 사례에서 '계속적인 과정'으로서의 인권영향평가를 강조하고 있다. 이를 위해 일정 주기마다 정기적으로 인권영향평가를 수행하도록 하고 있으며, 특히 인공지능 시스템이 적용되는 맥락이 중대하게 변경되었을 경우, 예를 들어 인공지능 시스템에 입력되는 데이터나 알고리즘이 변경되었을 때, 혹은 인공지능 시스템이 적용되는 범위가 확대되었을 때(예를 들어, 한 나라에 적용되는 인공지능 시스템이 다른 나라에 적용되는 경우), 인권영향평가를 다시 수행할 것을 권고하고 있다.

【참고】

• 유엔 <기업과 인권 이행지침>은 새로운 사업을 시작하기 전, 업무가 변화하는 중요한 결정을 내리기 전, 사업 주기 전반 등에 걸쳐 정기적으로 위험 및 영향 식별 및 평가를 반복하도록 한다.

• 덴마크 <디지털활동 인권영향평가>는 동일한 제품을 새로운(고위험) 시장에 출시할 때, 이용약관의 중대한 변경이 있을 때, 또는 특정 시장에서 제품을 철수하는 결정 등 디지털 사업, 제품 및 서비스의 규모, 범위, 사용 또는 적용이 변경될 때마다 인권 위험 및 영향이 재평가하도록 한다. 법률, 규정 또는 시장에 대한 중요한 변화가 있을 때, 또는 사회적, 정치적 상황에 중대한 변화가 있을 때에도 최초 인권영향평가의 결과를 재평가할 필요가 있다.

• 유럽평의회 <인권・민주주의・법치 영향평가>는 영향평가를 영향을 완화하기 위한 지속적인 학습 과정으로 고려할 필요성을 언급한다. 영향평가로 해당 인공지능을 변경했을 때, 부정적 영향이 완화되었는지 검토하기 위해 영향평가를 반복할 수 있다. 또한, 인공지능이 의도된 범위를 넘어 사용되거나 혹은 더 큰 시스템의 일부가 되는 것과 같이 새로운 인권 위험이 파악되었을 때도 영향평가를 반복할 수 있다.

• 캐나다 「자동화된 의사결정 훈령」은 자동화된 의사결정 시스템의 기능 또는 범위가 변경될 시 알고리즘영향평가를 갱신하도록 한다(6.1.3).

부록

인공지능 인권영향평가 1단계에서 평가팀은 여러 이해관계자로부터 자료와 의견을 수집하게 되는데, 인공지능 시스템 자체에 대한 정보를 파악하기 위해서는 해당 시스템을 기획, 개발, 설계, 디자인한 사람이나 팀(이하 개발팀)으로부터 관련 자료를 얻을 수밖에 없다. 혹은 외부의 인공지능 기술, 시스템, API 등을 활용할 경우에는 해당 업체로부터 관련 자료를 얻게될 것이다. 이를 위해 평가팀은 인권영향평가도구의 질의에 답을 하기 위해 개발팀 혹은 개발업체에 아래와 같은 질의를 통해 자료를 요청하거나 의견을 구할 수 있을 것이다. 물론 기획자, 개발자, 디자이너, 마케터 중 누가 이러한 정보를 갖고 있거나 의견을 제시할 수 있을지는 구체적인 맥락에 따라 달라질 것이다. 또한 아래 질의 중 일부는 다른 이해관계자에게도 적용될 수 있다. 평가팀은 개발자가 제공한 정보나 답변을 다른 이해관계자가 제공한 정보나 답변, 그리고 평가팀이 자체적으로 평가한 내용과 비교, 분석함으로써 객관적인 평가를 수행할 수 있다.

인공지능 인권영향평가 자료 수집 목적의 개발자에 대한 질의 목록

- 인공지능 시스템은 어떠한 문제를 해결하기 위한 것입니까, 즉 인공지능 시스템이 달성하고자 하는 목적 및 의도된 용도는 무엇입니까?

- 해당 인공지능 시스템의 기획, 개발, 설계, 디자인과 관련된 당사자는 어떻게 구성되어 있습니까?

- 해당 인공지능 시스템의 의도된 사용자는 누구입니까?

- 해당 인공지능 시스템의 사용으로 영향을 받는 사람이나 집단은 누구입니까?

- 해당 인공지능 시스템의 일부 기능을 외부 업체나 오픈소스에 의존하고 있습니까, 그렇다면 그 세부적인 내용은 무엇입니까?

- 데이터셋, 알고리즘 등 해당 인공지능 시스템과 관련된 정보(예를 들어, 데이터셋이나 알고리즘 등의 특성 및 이에 대한 평가, 외부업체의 제품을 구매할 경우 관련한 설명서, 사전학습 모델 가중치 등)를 제공해주십시오.

- 인공지능 시스템이 개인정보보호위원회 <인공지능(AI) 개인정보보호 자율점검표>의

모든 의무/권장 조항을 준수하고 있습니까?

• 해당하는 경우, 인공지능 시스템의 개인정보 처리에 대해 개인정보 영향평가를 수행하였습니까?

• 학습, 검증, 테스트 등 인공지능의 개발 과정에 사용되는 데이터셋에 대한 정보(예를 들어 데이터셋의 출처, 구조와 유형, 사전 처리 과정 등)를 제공해주십시오.

• 데이터셋의 정확성, 완전성, 최신성이 보장되고 있습니까, 그리고 이를 검토하기 위해 사용한 방법이 무엇입니까?

• 데이터셋이 인공지능 시스템이 사용될 맥락에 적합하도록 인구집단별 다양성과 대표성을 갖추었습니까?

• 데이터셋이 사상・신념, 건강, 인종이나 민족에 관한 정보, 생체인식정보 등 민감정보를 포함하고 있습니까?

• 데이터셋과 관련된 앞의 정보를 확인하는 것이 불가능하거나 불필요할 경우, 그 이유는 무엇입니까? 그리고 이 경우 데이터셋의 편향성을 방지할 수 있는 다른 방안은 무엇입니까?

• 인공지능 시스템 외의 다른 대안 혹은 채택된 알고리즘(또는 사전학습 모델 가중치) 외에 다른 대안에 대해 검토한 바가 있습니까?

• 인공지능 시스템에 사용된 알고리즘(또는 사전학습 모델 가중치)이 목적 달성에 적합한 이유는 무엇입니까?

• 인공지능 시스템이 의도한대로 작동하는지 성능을 측정하기 위한 지표와 방법은 무엇이니까?

• 정확도와 오류율 등 성능의 수준은 어떻게 설정되었으며, 그것을 측정하는 방법은 무엇입니까?

• 해당 인공지능 시스템이 특정한 결정(출력)을 내리는데 관련된 요소를 추적할 수 있도록 관련된 정보(예를 들어, 결정의 내역이나 시스템에 대한 모든 변경사항 등에 대한 로그기록)가 기록되고 있습니까?

• 해당 인공지능 시스템이 특정한 결정(출력)을 내린 이유나 근거에 대해 사용자 혹은 영향을 받는 이해관계자에게 쉽게 설명할 수 있습니까?

• 인공지능 시스템의 사용자 혹은 상호작용하는 사람에게 상대방이 사람이 아니라 인공지능 시스템이라는 사실, 혹은 자신이 받은 결과물이나 결정이 인공지능 시스템에 의한 것이라는 점을 적절한 방법으로 알릴 수 있습니까?

• 인공지능 시스템의 영향을 받는 당사자가 인지하지 못하는 방식으로 인공지능 시스템이 작동하지 않도록 당사자에게 적절하게 알릴 수 있습니까?

• 인공지능 시스템의 결과물에 기반한 결정에서 인간의 역할과 인간이 재량권을 갖고 개입할 수 있습니까?

• 인공지능 시스템이 의도한대로 작동하지 않을 경우, 인공지능 시스템의 운영자 혹은 사용자가 언제든지 시스템을 정지시킬 수 있습니까?

• 인공지능 시스템 보안에 대한 가능한 위협이 무엇이고 보안이 침해되었을 경우 발생할 수 있는 결과 혹은 해악은 무엇입니까?

• 인공지능 시스템의 학습 및 테스트에 활용되는 데이터셋에 대해 취해진 보안조치는 무엇입니까?

• 인공지능 시스템의 전체 수명주기 동안 발생할 수 있는 잠재적인 공격에 대비하여 무결성, 가용성, 기밀성, 견고성 등 보안에 요구되는 요소를 보장하기 위한 보안조치는 무엇입니까?

• 해당 인공지능 시스템은 언어, 나이, 장애, 신체적 조건 등에 상관없이 누구나 사용할 수 있도록 설계되었습니까, 그렇지 않다면 그 이유는 무엇입니까?

• 인공지능 시스템의 전체 혹은 일부 소프트웨어를 외부에서 개발된 것을 사용할 경우, 필요한 경우 알고리즘 혹은 소스코드 등 소프트웨어를 적절하게 수정, 변경할 수 있는 권한을 부여받았습니까?

• 인공지능 시스템이 오류로 인하여 의도하지 않은 방식으로 작동할 경우 어떠한 부정적 결과를 야기할 수 있습니까?

• 인공지능 시스템이 의도적으로 악용될 가능성이 있습니까. 그럴 경우 어떠한 부정적인 결과를 야기할 수 있습니까?

제6장 결론 및 정책권고

인공지능 기술은 인간과 인권에 피할 수 없는 위험을 야기한다. 인공지능 기술은 기존 기계문명의 혁신을 뛰어넘어 인간의 본질적 영역으로 여겨왔던 내면의 의사결정까지 수행하는 수준에 이르고 있다. 또한 인공지능은 인간이 알지 못하는 과정을 수행하여 결과를 산출하고 있다. 인공지능은 인권의 문제를 새로운 방식으로 제기하고 있다.

인공지능 및 알고리즘에 대한 영향평가가 주목을 받는 이유는 인공지능 기술에 잠재하는 인간과 인권에 대한 위험성을 예방하고 완화하기 위한 하나의 제도적 방안으로 평가받고 있기 때문이다. 최근 유럽연합, 캐나다, 영국, 미국 등 주요 국가들은 인공지능 및 알고리즘에 대한 영향평가 제도를 도입해 왔다. 특히 세계 각국은 공공부문 인공지능과 민간부문 고위험 인공지능이 사람들에게 미치는 부정적 영향에 주목하고 이를 예방적으로 해결하기 위한 방안으로 다양한 인공지능 및 알고리즘 평가를 제안하고 있다.

인공지능 영향평가 제도는 크게 위험기반 접근과 인권기반 접근에 따라 구분할 수 있는데, 본 연구는 인공지능에 대한 위험영향평가로 분류되는 유럽연합, 캐나다, 영국, 미국의 사례와 함께, 유엔 인권규범, 유럽평의회, 덴마크, 네덜란드에서 제안하였거나 도입 중인 인공지능 인권영향평가제도의 기준을 검토하여 인공지능 인권영향평가도구(안)을 개발하였다.

하지만 인공지능 기술은 지금도 비약적으로 발전하고 있고 그 속도와 방향도 예측불가능하다. 인권영향평가의 구체적인 준칙과 방법을 정립하고 시행하는 데도 여러 가지 혼란이 있을 수밖에 없다. 신기술 도입에 따른 여러 기업과 규제당국의 이해관계가 얽혀 있기도 하고, 기존 규제법제나 영향평가와의 제도적 관계나 정합성 여부도 함께 검토되어야 한다.

최근 기업의 ESG 이슈가 각광을 받으면서 인권실사에 대한 관심도 늘어나고 있다. 인권영향평가가 인권실사의 핵심도구라는 점에서 민간기업이 인권영향평가를 사실상 경영평가의 핵심요소로 활용할 가능성이 매우 높다. 그러나 이른바 인권과 같은 비재무적 요소가 기업가치에 어떤 영향을 미치는지에 대해 주목하는 인권 위험관리와 이해관계자의 인권에 어떤 영향을 미치는지에 주목하는 인권영향평가는 구분되어야 한다. 인공지능 인권영향평가는 기업가치에 대한 위험성을 평가하는 도구가 아니라 영향을 받는 당사자의

인권에 대한 위험성을 평가하는 도구로 이해되어야 한다.

또한 인공지능 인권영향평가는 인공지능으로 인한 침해의 심각도를 고려한다는 점에서 위험영향평가의 요소를 반영하고 있지만, 특정 인권 항목이 아니라 모든 인권 항목에 대하여 전체적, 보편적, 포괄적 접근방식을 강조하고 영향을 받는 당사자의 참여와 이들에 대한 투명성을 중시하는 국제인권규범을 기준으로 삼는다는 점에서 다른 영향평가와 다르다는 점에도 유념해야 한다.

유럽평의회는 알고리즘 인권영향에 대한 각료위원회의 권고를 채택하면서, 특히 인권영향평가를 국가와 민간이 의무적으로 취하여야 할 예방적 조치로 보고 상세한 요구사항을 설명한 바 있다. 유엔 인권최고대표도 국가와 기업에 대하여 인공지능 시스템의 설계, 개발, 배치, 판매, 구입, 운영의 수명 주기 전반에 걸쳐 체계적으로 인권실사를 수행할 것을 권고하고, 그 인권실사의 핵심 요소는 정례적이고 포괄적인 인권영향평가여야 한다고 강조하였다.

이에 인공지능의 개발과 활용에서 인간과 사회, 환경에 대한 부정적 영향을 방지하고 인간의 존엄성과 인권 보장을 위해 다음과 같은 인공지능 인권영향평가도구(안)를 제시한다. 이 도구(안)는 인공지능 관련 기준 및 관련 법령에 적용할 수 있도록 구체적인 인권영향평가 도구로 고안되었다.

인공지능 인권영향평가도구(안)

1) 인공지능 인권영향평가 개요

가) 인공지능 인권영향평가의 대상 : 고위험 인공지능

법률에서 금지하는 인권 침해 또는 차별 대우를 목적으로 하거나 법률에서 금지하는 개인정보의 처리를 목적으로 하는 인공지능 등 위험의 완화 내지 제거가 불가능한 인공지능은 일응 ‘금지대상 인공지능’에 해당하여 인권영향평가의 대상에서 제외된다.

인공지능 인권영향평가가 대상인 인공지능은 공공기관이 직접 개발하거나 조달하는 모든 인공지능 및 민간에서 활용하는 인권 침해의 위험성이 높은 인공지능, 즉 고위험 인공지능이다. 물론 금지되는 인공지능의 기준과 마찬가지로 현 단계에서는 무엇이 고위험 인공지능인지에 대한 명확한 사회적인 합의가 존재하지 않으며, 고위험 인공지능에

대한 인권영향평가의 실시를 의무화하는 법률도 존재하지 않는다. 고위험 인공지능에 대한 인권영향평가 실시 의무화를 위해서는 고위험 인공지능의 범위 및 인권영향평가 수행 의무화를 내용으로 하는 입법화가 선행될 필요가 있다.

그러나 인권영향평가는 인공지능의 잠재적 위험을 체계적으로 검토하고 이해관계자와의 대화와 협력을 통해 사전에 방지, 완화하자는 취지로 시행된다. 따라서 국내외에서 고위험 인공지능이라고 거론되는 인공지능의 경우 본 인권영향평가를 수행할 것이 강하게 권고된다.

- 항공, 자동차, 철도, 기계, 장난감의 안전 관련 구성요소이거나, 승강기, 무선 장비 및 의료 기기 등의 안전 관련 구성요소 또는 제품 그 자체인 경우
- 사람의 생체정보를 활용하여 신원확인을 수행하는 경우
- 교통, 수도, 가스, 전기 등 중요 사회기반시설의 관리·운용에 활용하는 경우
- 소방, 응급의료 등 필수 공공·민간 서비스에 활용하는 경우
- 채용, 인사평가 또는 직무 배치의 결정에 사용하는 경우
- 공공 지원 혜택의 자격 및 수혜 적격성을 평가하기 위하여 사용하는 경우
- 범죄의 수사, 공소의 제기 및 유지, 형 및 보안처분의 집행에 사용하는 경우
- 이주, 망명 및 출입국 관리에 활용하는 경우
- 사실의 인정 및 법률 해석, 적용 등 법관의 업무를 지원하는 데 사용하는 경우
- 군 또는 정보기관에서 사용하는 경우

나) 인공지능 인권영향평가 시기

공공기관의 경우 위험도와 무관하게, 민간의 경우 고위험 인공지능을 개발하거나, 고위험인공지능을 사업 또는 정책의 기반기술로 도입하기 이전에 인권영향평가를 수행하되, 사전영향평가에만 한정하지 않고, 정기적, 사후적 평가를 통한 지속적인 관리와 모니터링을 전제한다. 고위험인공지능에 비하여 위험의 정도가 덜한 인공지능의 경우 국가인권위원회의 직권 지정 또는 이해관계자의 요청에 따른 검토를 거쳐 국가인권위원회의 지정에 따라 인권영향평가를 수행할 수 있고, 개발 또는 도입 주체의 자발적인 요구에 의

해서도 수행될 수 있다. 사전 영향평가의 시점은 인공지능기술 개발 또는 도입에 관한 구상이 구체화된 시점이다.

다) 인공지능 인권영향평가 수행 주체

구체적인 영향평가의 수행은 인공지능의 개발 주체 및 관련된 사업부서와는 독립된 별도의 조직(예를 들어, 인공지능 윤리, 인권 경영, ESG 경영 등을 담당하는 부서) 또는 독립성과 인권 분야에 대한 전문성 및 인공지능 기술에 대한 전문성을 갖춘 제3의 기관이 수행하도록 한다.

라) 인공지능 인권영향평가의 절차

본 연구는 인공지능 인권영향평가의 이행단계를 4단계로 범주화하였다.

- 1단계 : 계획 및 준비
- 2단계 : 분석 및 평가
- 3단계 : 개선 및 구제
- 4단계 : 공개 및 점검
- 공통 : 이해관계자의 참여

한편, 영향평가에 따른 결과서는 국가인권위원회에 제출되며, 국가인원위원회는 영향평가결과를 검토한 후 미흡한 점에 대한 개선을 권고하거나, 위험성에 대한 완화 조치 또는 제거 조치가 불가능하다고 판단하는 경우 개발 또는 활용의 중단을 권고하는 등 의견을 제시할 수 있도록 한다.

아래는 인권영향평가 과정에서 점검해야 할 항목을 체크리스트 방식으로 제시한다.

2) 인공지능 인권영향평가도구

【1단계 : 계획 및 준비】

가. 인권영향평가 계획

Q1-1-1. 인권영향평가의 대상이 되는 인공지능 시스템 혹은 프로젝트는 무엇입니까.
인공지능 시스템 혹은 프로젝트명 :

Q1-1-2. 인권영향평가를 수행하는 책임자는 누구입니까.
책임자의 성명과 소속 :

Q1-1-3. 평가팀은 인권영향평가를 수행하기에 충분한, 인공지능 기술 및 인권에 대한 전문성을 갖추고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
평가팀의 구성, 팀원의 역할, 전문분야 등을 설명하시오.
설명 ( )

Q1-1-4. 조직 내에 인권영향평가 수행의 요건, 주체, 절차 등을 상세히 규정한 정책을 두고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-1-5. 인권영향평가를 내실있게 수행하는데 충분한 인적, 재정적 자원이 확보되어 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-1-6. 인권영향평가 결과의 수용 여부를 결정할 수 있는 조직 내 최종 책임자 혹은 책임단위에 인권영향평가 결과보고서를 보고하는 절차가 명확하게 규정되어 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-1-7. 인권영향평가를 내실있게 수행할 수 있도록, 평가팀이 평가 대상이 되는 인공지능 시스템의 개발 혹은 활용과 관련한 부서 및 담당자에게 협조를 요청하고, 인권영향평가에 필요한 핵심 자료에 접근할 수 있는 권한이 보장되어 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q1-1-8. 인공지능 시스템은 어떠한 문제를 해결하기 위한 것입니까, 즉 인공지능 시스템이 달성하고자 하는 목적 및 의도된 용도는 무엇입니까.
인공지능 시스템의 목적 :

Q1-1-9. 해당 인공지능 시스템이 적용되는 분야에서 인공지능 시스템의 기능, 요건, 제한 등에 영향을 미치는, 인권 보호를 위해 요구하고 있는 법령상의 요건(법률, 시행령, 시행규칙 등의 관련 조항)이 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
만일 있다면, 해당 법률, 시행령, 시행규칙 등의 관련 조항은 무엇입니까.
설명 ( )

Q1-1-10. 인공지능 시스템이 인권에 미치는 영향을 평가하기 위하여 조직 내외부의 다양한 이해관계자의 의견을 검토할 필요가 있습니다. 다양한 이해관계자를 인권영향평가 과정에 참여시키기 위해서는 우선 누가 이해관계자인지 파악해야 합니다. 아래 질의에서 이해관계자가 누구인지 가능한 구체적으로 적어주세요.

Q1-1-10-1. 해당 인공지능 시스템에 대한 공정한 인권영향평가를 위해, 조직 내부에서 해당 인공지능 시스템의 개발 및 운영에 관련된 이해관계자(예를 들어, 기획, 개발, 디자인, 유지보수, 정책, 데이터 거버넌스, 영업 등 담당 부서)의 참여가 중요합니다. 이를 위해 조직 내부에서 참여할 수 있는 이해관계자는 누구입니까.
설명 ( )

Q1-1-10-2. 해당 인공지능 시스템에 대한 공정한 인권영향평가를 위해, 조직 외부에서 해당 인공지능 시스템의 개발 및 운영에 관련된 이해관계자(예를 들어, 외부 개발업체, 위탁업체, 유지보수업체, 감독기구, 전문가 집단 등)의 참여 역시 중요합니다. 이를 위해 조직 외부에서 참여할 수 있는 이해관계자는 누구입니까.
설명 ( )

Q1-1-10-3. 인공지능 시스템의 사용자는 누구입니까.

설명 ( )

Q1-1-10-4. 인공지능 시스템의 사용으로 영향을 받는 사람이나 집단은 누구입니까.
설명 ( )

Q1-1-10-5. 인공지능 시스템의 사용으로 영향을 받는 개인이나 집단에 아동, 노인, 장애인, 여성, 외국인, 성소수자, 저학력자, 비정규직 노동자, 경제적 약자, 낙후지역 등 취약하거나 소외된 집단이 포함되어 있다면 구체적으로 적어주세요.
설명 ( )

나. 조사

Q1-2-1. 인공지능 시스템이 인권에 미치는 영향을 이해하기 위해서는 해당 시스템에 대한 이해가 필요합니다. 데이터셋, 알고리즘 등 해당 인공지능 시스템과 관련된 정보(예를 들어, 데이터셋이나 알고리즘 등의 특성 및 이에 대한 평가, 외부업체의 제품을 구매할 경우 관련한 설명서, 사전학습 모델 가중치 등)를 확보하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-2-2. 인공지능 시스템이 도입, 활용될 분야 혹은 시공간적인 특성 및 맥락과 관련된, 인권에 영향을 미칠 수 있는 요소에 대한 자료를 확보하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-2-3. 앞서 파악한, 인공지능 시스템의 이해관계자로부터 해당 시스템이 인권에 미칠 영향에 대한 의견을 수렴하거나 협의하고 이를 문서화 하였습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q1-2-4. 이해관계자 의견을 수렴하거나 협의할 때 다음과 같은 내용을 포함합니까.
- 협의한 이해관계자의 성명, 소속, 연락처
- 협의한 일자

- 인공지능 시스템에 대해 이해관계자에게 제공한 자료
- 인공지능 시스템에 대한 이해관계자의 의견

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q1-2-5. 인공지능 시스템의 활용으로부터 영향을 받는 이해관계자, 특히 취약하거나 소외된 집단과의 협의를 포함하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q1-2-6. 관련 자료를 수집하거나 이해관계자의 의견을 수렴할 때 자료의 기밀성을 유지하고 이해관계자의 개인정보를 보호할 수 있는 조치를 취하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

【2단계 : 분석 및 평가】

가. 인공지능 기술과 관련된 영향 분석 및 평가

(1) 개인정보보호

Q2-1-1. 인공지능 시스템이 개인정보보호위원회 <인공지능(AI) 개인정보보호 자율점검표>의 모든 의무/권장 조항을 준수하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-2. 해당 인공지능 시스템의 개발 혹은 운영 과정의 개인정보 처리가 개인정보 보호법 상 개인정보 영향평가를 의무적으로 수행해야 하는 경우, 개인정보 영향평가를 수행하였는지 확인하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

(2) 데이터

Q2-1-3. 학습, 검증, 테스트 등 인공지능의 개발 과정에 사용되는 데이터셋에 대한 정보, 예를 들어 데이터셋의 출처, 구조와 유형, 사전 처리 과정 등에 대한 정보를 확보하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-4. 데이터셋의 정확성, 완전성, 최신성을 확인하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

이를 검토하기 위해 사용한 방법은 무엇입니까.

설명 ( )

Q2-1-5. 데이터셋이 인공지능 시스템이 사용될 맥락에 적합하도록 인구집단별 다양성과 대표성을 갖추었는지 확인하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

이를 검토하기 위해 사용한 방법은 무엇입니까.

설명 ( )

Q2-1-6. 데이터셋이 사상·신념, 건강, 인종이나 민족에 관한 정보, 생체인식정보 등 민감정보를 포함하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-7. 대리 변수를 통해 민감정보의 추정이 가능한지 여부를 검토하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-8. Q2-1-3 ~ Q2-1-7의 질의 전체 혹은 일부에 대한 확인이 불가능하거나 이를 확인하는 것이 불필요하다고 판단하는 경우, 그 이유는 무엇입니까. 또한 그러한 경우 데이터셋의 편향성을 방지할 수 있는 다른 방안은 무엇입니까.

설명 ( )

(3) 알고리즘의 성능과 신뢰성

Q2-1-9. 인공지능 시스템 외의 다른 대안 혹은 채택된 알고리즘(또는 사전학습 모델 가중치) 외에 다른 대안에 대한 검토가 있었습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
인공지능 시스템에 사용된 알고리즘(또는 사전학습 모델 가중치)이 목적 달성에 적합한 이유는 무엇입니까.
설명 ( )

Q2-1-10. 인공지능 시스템이 의도한대로 작동하는지 성능을 측정하기 위한 지표와 방법을 갖고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q2-1-11. 정확도와 오류율 등 성능의 수준은 의도한 목적에 적합한 정도로 설정되었습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
인공지능 시스템의 정확도와 오류율 등 성능은 어떻게 측정합니까.
설명 ( )

(4) 차별금지

Q2-1-12. 인공지능 시스템이 활용 과정에서 합리적인 이유없이 인종, 종교, 장애, 나이, 학력, 직업, 출신 지역, 언어, 정치성향, 신체조건, 외모, 피부색, 병력, 성별, 성적 지향, 사회적 신분, 경제적 지위 등 개인과 집단의 특성에 따라 특정 집단에 대한 차별을 야기하거나 혹은 기존의 차별을 악화할 가능성이 있는지 검토하였습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q2-1-13. 인공지능 시스템 개발 과정에서 알고리즘에 의한 구조적 차별을 사전에 방지

하기 위하여, 기획, 개발, 디자인, 마케팅, 경영진 등 조직 구성원의 다양성 확보, 구성원에 대한 반차별 교육, 조직 내 인공지능 윤리 정책 수립 등의 대책을 마련하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

(5) 설명가능성과 투명성

Q2-1-14. 해당 인공지능 시스템이 특정한 결정(출력)을 내리는데 관련된 요소를 추적할 수 있도록 관련된 정보(예를 들어, 결정의 내역이나 시스템에 대한 모든 변경사항 등에 대한 로그기록)가 기록되고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-15. 해당 인공지능 시스템이 특정한 결정(출력)을 내린 이유나 근거에 대해 사용자 혹은 영향을 받는 이해관계자에게 설명할 수 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-16. 설명가능성 여부가 인권에 영향을 미칠 경우, 해당 인공지능 시스템의 작동이나 특정한 결정의 근거에 대해 기술 전문가가 아닌 이해관계자가 충분히 이해할 수 있는 방식으로 설명할 수 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-17. 정확도와 오류율 등 인공지능 시스템의 성능, 어떤 결정을 내리는데 사용되는 매개변수 및 가중치, 적절한 사용법, 장점과 한계 등에 대해 사용자가 이해할 수 있는 방식으로 충분한 정보를 제공하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q2-1-18. 인공지능 시스템의 소스코드가 공개되거나 이를 요구하는 이해관계자에게 제공될 수 있습니까. 소스코드가 제공될 수 있다면, 누구에게 어떤 조건으로 제공됩니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

(6) 자동화 정도와 인간의 개입

Q2-1-19. 사람과 상호작용하는 인공지능 시스템의 경우, 인공지능 시스템의 사용자 혹은 상호작용하는 사람에게 상대방이 사람이 아니라 인공지능 시스템이라는 사실, 혹은 자신이 받은 결과물이나 결정이 인공지능 시스템에 의한 것이라는 점을 적절하게 알릴 수 있는 조치를 취하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q2-1-20. 인공지능 시스템이 영향을 받는 이해관계자가 인지할 수 없도록 은밀하게 작동할 수 있는 경우(예를 들어, 원격에서 작동하는 얼굴인식 시스템이 대상자 모르게 얼굴인식을 통해 신원을 파악하는 경우) 영향을 받는 당사자가 인지하지 못하는 방식으로 인공지능 시스템이 작동하지 않도록 당사자에게 적절하게 알릴 수 있는 조치를 취하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q2-1-21. 인공지능 시스템의 결과물에 기반한 결정에서 인간의 역할과 인간이 재량권을 갖고 개입할 수 있는 범위와 절차가 정의되어 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q2-1-22. 인공지능 시스템이 의도한대로 작동하지 않을 경우, 인공지능 시스템의 운영자 혹은 사용자는 언제든지 시스템을 정지시킬 수 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

(7) 보안

Q2-1-23. 인공지능 시스템의 특성과 활용되는 분야 등을 고려했을 때, 인공지능 시스템 보안에 대한 가능한 위협이 무엇이고 보안이 침해되었을 경우 발생할 수 있는 결과 혹은 해악은 무엇입니까.
설명 ( )

Q2-1-24. 인공지능 시스템의 학습 및 테스트에 활용되는 데이터셋에 대해 충분한 안전조치가 적용되었습니까. 데이터 오염 등 데이터에 대한 다양한 유형의 공격에 대한 대응이 고려되었습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q2-1-25. 인공지능 시스템의 전체 수명주기 동안 발생할 수 있는 잠재적인 공격에 대비하여 무결성, 가용성, 기밀성, 견고성 등 보안에 요구되는 요소를 보장하기 위한 조치를 취했습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

(8) 접근성

Q2-1-26. 인공지능 시스템이 언어, 나이, 장애, 신체적 조건 등에 상관없이 누구나 사용할 수 있도록 인공지능 시스템의 인터페이스가 보편적 설계 원칙에 따라 설계되었습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q2-1-27. 보편적 설계 원칙에 따라 설계하지 않은 합리적인 이유가 있다면 그것은 무엇입니까.
설명 ( )

(9) 라이선스

Q2-1-28. 인공지능 시스템의 전체 혹은 일부 소프트웨어를 외부에서 개발된 것을 사용할 경우, 인공지능 시스템에 의한 잠재적 인권 침해를 방지, 완화하기 위하여 필요한 경우 알고리즘 혹은 소스코드 등 소프트웨어를 적절하게 수정, 변경할 수 있는 권한에 대해 외부 개발업체와 명확한 합의가 이루어져 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

나. 인권에 미치는 영향 및 심각도

(1) 영향을 받는 인권

Q2-2-1. 인공지능 시스템이 도입, 활용될 경우 시민들의 인권에 미칠 수 있는 부정적인 영향 혹은 위험은 무엇입니까. 누구의 인권이 어떤 방식으로 침해될 수 있습니까.
설명 ( )

Q2-2-2. 인공지능 시스템이 오류로 인하여 의도하지 않은 방식으로 작동할 경우 나타날 수 있는 부정적인 결과에 대해 검토한 바 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
의도하지 않은 방식으로 작동할 경우 침해되는 인권은 무엇입니까
설명 ( )

Q2-2-3. 인공지능 시스템이 의도적으로 악용될 가능성이 있습니까. 어떠한 방식으로 오용될 수 있는지에 대해 검토하였습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
악용될 경우 나타날 수 있는 부정적인 결과, 혹은 침해되는 인권은 무엇입니까
설명 ( )

(2) 인권에 미치는 영향의 심각도

Q2-2-4. 인공지능 시스템이 인권에 미치는 부정적 영향의 범위가 어떠합니까. 전체 인

구 혹은 어떠한 특정 집단에 대하여 어느 정도의 범위(대, 중, 소)로 영향을 미칠 수 있습니까. (부정적 영향을 받을 수 있는 인권이 여러 개인 경우 각각에 대해서 평가가 필요함. 아래 질의에 대해서도 동일함)
설명 ( )

Q2-2-5. 인공지능 시스템이 인간의 생명, 건강, 안전, 인권, 기본적 삶 등에 미치는 부정적 영향의 규모 혹은 크기가 어떠합니까. (대, 중, 소)
설명 ( )

Q2-2-6. 인공지능 시스템이 인권에 미치는 부정적 영향이 사후에 구제나 회복이 어느 정도 가능합니까. (완전히 회복 가능, 부분적으로 회복 가능, 회복 불가능)
설명 ( )

Q2-2-7. 인공지능 시스템이 인권에 미치는 부정적 영향이 여러 개일 경우 서로 상충하는 인권이 있습니까. 또한 그 심각성 때문에 우선적인 대응이 필요한 인권은 무엇입니까.
설명 ( )

【3단계 : 개선 및 구제】

가. 방지

Q3-1-1. 데이터에 대한 개선, 알고리즘의 수정, 시스템 설계 변경 등 2단계(영향 분석 및 평가)에서 파악된 중대한 인권 위험을 방지하기 위해 어떠한 조치를 취했습니까.
설명 ( )

나. 완화

Q3-2-1. 2단계(분석 및 평가)에서 파악된 인권 위험을 완전히 방지하기 곤란한 경우, 위험을 완화하기 위해 어떠한 조치를 취했습니까.
설명 ( )

Q3-2-2. 인공지능 시스템의 잔존하는 위험성에 대해 사용자 및 영향을 받는 이해관계자에게 충분한 정보를 제공하고 올바른 작동 방법에 대해 적절한 교육을 제공하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q3-2-3. 인공지능 시스템의 인권 침해 위험이 클 수 있는 특정한 사용을 허용하지 않도록 이용약관이나 여타의 집행체계에서 금지하는 절차를 취하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q3-2-4. 2단계(분석 및 평가)에서 파악된 중대한 인권 위험이 완화되지 않고 남아있을 경우 그 이유를 문서화하고 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q3-2-5. 중대한 위험에 대한 방지 및 완화 조치를 취하기 힘들거나, 이러한 조치를 취해도 여전히 중대한 위험이 남아있을 경우 인공지능 시스템의 개발 및 활용을 중단할 계획을 갖고 있습니까

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

다. 구제

Q3-3-1. 인공지능 시스템의 결정에 의해 영향을 받는 사람이 인공지능 시스템의 결정에 이의를 제기하거나 침해된 권리의 구제를 요구할 수 있는 절차를 마련하고, 이에 대한 정보를 누구나 쉽게 접근할 수 있도록 일반에 공개하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q3-3-2. 인공지능 시스템에 의해 영향을 받는 사람들에게 인공지능 시스템의 사용을

거부할 수 있는 선택권(옵트아웃 권리)을 제공하거나 이의를 제기할 수 있는 수단이 마련되어 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q3-3-3. 인공지능 시스템이 내린 결정에 의해 영향을 받는 이해관계자가 인공지능 시스템의 적용을 거부할 경우, 인간의 지원 혹은 인공지능이 아닌 시스템의 적용을 대안으로 제시할 것을 고려하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q3-3-4. 인공지능 시스템의 결정에 대한 이의제기나 권리구제 요구가 정당할 경우, 문제의 의사결정을 번복하거나 권리를 복구하거나 손해배상을 할 수 있는 절차가 마련되어 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

라. 이해관계자와의 의견수렴 및 협의

Q3-4-1. 인공지능 시스템의 인권 위험을 방지, 완화하고 인권을 침해받은 사람의 권리를 구제하기 위한 조치에 대해서 관련 이해관계자(1단계에서 파악한 이해관계자)의 의견을 수렴하거나 협의를 진행하였습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

Q3-4-2. 공공기관이 도입하는 인공지능 시스템의 경우, 가능한 모든 이해관계자가 참여할 수 있도록 의견 수렴을 위한 공청회를 실시하고 있습니까.

☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

【4단계 : 공개 및 점검】

가. 인공지능 시스템의 주요 요소의 공개

Q4-1-1. 인공지능 시스템이 사용하는 데이터와 알고리즘 등의 주요 요소를 일반에 공개하고 이해할 수 있는 방식으로 쉽게 설명하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

나. 인권영향평가 결과 공개

Q4-2-1. 인권영향평가 보고서 전체 혹은 주요 내용을 일반에 공개합니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q4-2-2. 인권영향평가 보고서를 감독기구인 국가인권위원회에 제공하고, 효과와 한계에 대해 협의하는 절차가 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

다. 인공지능 시스템에 대한 모니터링

Q4-3-1. 인공지능 시스템이 도입되거나 운영이 시작된 후에 그 성능과 인권에 미치는 부정적 영향, 완화 조치 및 구제 정책의 효과성을 확인하기 위해, 인공지능 시스템의 수행을 모니터링하고 기록에 남길 수 있는 메커니즘을 수립하였습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q4-3-2. 인공지능 시스템이 의도한 대로 작동하지 않거나 인권에 미치는 부정적인 영향이 확인되었을 때, 관련된 책임을 명확히 하고 인공지능 시스템을 개선하며 부정적 영향을 완화하기 위해 필요한 절차를 수립하였습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음

설명 ( )

라. 인권영향평가에 대한 점검

Q4-4-1. 인권영향평가 수행의 효과와 한계를 점검하고, 개선할 수 있는 절차를 마련하였습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

마. 인권영향평가의 재수행

Q4-5-1. 인공지능 시스템에 대해 정기적으로 (예를 들어 1년) 인권영향평가를 수행하는 절차를 마련하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

Q4-5-2. 인공지능 시스템의 핵심적인 기능이 변경되거나 환경적 요인 혹은 적용 범위가 변경되었을 경우, 인권영향평가를 다시 수행하는 절차를 마련하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
설명 ( )

한편, 인공지능 인권영향평가 1단계에서 평가팀은 여러 이해관계자로부터 자료와 의견을 수집하게 되는데, 인공지능 시스템 자체에 대한 정보를 파악하기 위해서는 해당 시스템을 기획, 개발, 설계, 디자인한 사람이나 팀으로부터 관련 자료를 얻을 수밖에 없다. 외부의 인공지능 기술, 시스템, API 등을 활용할 경우에는 해당 업체로부터 관련 자료를 얻게 될 것이다. 평가팀은 인권영향평가도구의 질의에 답을 하기 위해 개발팀 혹은 개발 업체에 자료를 요청하거나 의견을 구할 수 있다. 평가팀은 개발자가 제공한 정보나 답변을 다른 이해관계자가 제공한 정보나 답변, 그리고 평가팀이 자체적으로 평가한 내용과 비교, 분석함으로써 객관적인 평가를 수행할 수 있다.

참 고 문 헌

과학기술정보통신부(2022). 2022 신뢰할 수 있는 인공지능 개발 안내서(안).

광주광역시(2022). 2021년 시정백서.

개인정보보호위원회(2021). 인공지능(AI) 개인정보보호 자율점검표.

국가인권위원회(2011). 기업과 인권 이행지침: 유엔 ‘보호, 존중, 구제’ 프레임워크의 실행.

국가인권위원회(2014). 인권영향평가 및 관리에 관한 지침(HRIAM 가이드).

국가인권위원회(2018). 공공기관 인권경영 매뉴얼.

국가인권위원회(2022). 인공지능 개발과 활용에 관한 인권 가이드라인.

금융위원회(2022). 금융분야 AI 개발·활용 안내서.

김기중, 오정미, 오철우, 장여경, 전치형, 김민(2021). 인공지능(AI) 개발과 활용에서의 인권 가이드라인 연구. 사단법인 정보인권연구소. 국가인권위원회 연구용역보고서.

김동현(2022). 의무적 인권실사의 해외 입법 동향과 국내 법제화 방안. 서강법률논총 제11권 제1호, p107-149.

김종철, 강은지, 김동현, 김두나, 김진, 나현필, 박예안, 정신영, 한정민(2020). 공공기관·공기업 인권영향평가 현황 실태조사 및 개선방안 연구. 사단법인 어필. 국가인권위원회 연구용역보고서.

박준석(2022). 인권영향평가(HRIAs)의 현실과 과제: 자치법규 인권영향평가를 중심으로. 법학연구 제69집.

법무부(2021). 기업과 인권 길라잡이.

서울특별시교육청(2021). 인공지능(AI) 공공성 확보를 위한 현장 가이드라인.

안국진(2018). 수원시 공공건축물 인권영향평가 실행방안 연구, 수원시정연구원.

이상수(2015). 인권실사의 개념과 법제도화 가능성, 법과 기업연구 제5권 제1호. 71-103면.

이준일, 김지혜, 이승택, 박진아, 남중권(2015). 경찰인권영향평가제도 수립을 위한 연구용역보고서: 경찰 인권영향평가 제도화 방안을 중심으로. 고려대학교 산학협력단. 경찰청 연구용역보고서.

이준일, 유승익, 이승택, 박진아(2018). 공공부문 인권영향평가 도입 방안 실태조사. 고려대학교 산학협력단. 국가인권위원회 연구용역보고서.

이충은, 노진석(2018). 인권영향평가의 제도화 방안에 관한 연구. 법과 정책 제24집 제2호. 제주대학교 법과정책연구소.

정영선(2013). 지방자치단체의 인권제도 발전 방향과 과제. 법학논집 제18권 제2호. 85-118면.

최유(2015). 인권영향평가에 관한 연구. 입법평가연구 제9호. 423-456면.

한국도시연구소(2019). 국토연구원 인권영향평가 연구. 국토연구원 연구용역보고서.

한국환경공단(2022). 2021년 한국환경공단 인권영향평가 결과보고서.

Ada Lovelace Institute(2020). Algorithmic impact assessment: AIA template. <https://www.adalovelaceinstitute.org/resource/aia-template/ (접근일: 2022. 8. 15)>.

Ad hoc Committee on Artificial Intelligence(2020). Human Rights, Democracy and Rule of Law Impact Assessment of AI systems. CAHAI-PDG(2021)05. <https://rm.coe.int/cahai-pdg-2021-05-2768-0229-3507-v-1/1680a291a3 (접근일: 2022. 8. 15)>.

BSR(2018). Human Rights Impact Assessment - Facebook in Myanmar. <https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf (접근일: 2022. 11. 1)>.

BSR(2019). Google Celebrity Recognition API Human Rights Assessment - Executive Summary. <https://www.bsr.org/reports/BSR-Google-CR-API-HRIA-Executive-Summary.pdf (접근일: 2022. 11. 1)>.

Council of Europe(2020). Guidelines on addressing the human rights impacts of algorithmic systems. Appendix to Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems.

Council of Europe Commissioner for Human Rights(2019). Unboxing artificial intelligence: 10 steps to protect human rights. <https://www.coe.int/en/web/commissioner/-/unboxing-artificial-intelligence-10-steps-to-protect-human-rights (접근일: 2022. 8. 15)>.

Désirée Abrahams (IBLF), Yann Wyss (IFC) (2010). Guide to Human Rights Impact Assessment and Management, HRIAM. <https://www.unglobalcompact.org/library/25 (접근일: 2022. 12. 1)>. 번역본: 국가인권위원회(2014).

European Commission(2020a). WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust. <https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en (접근일: 2022. 11. 1)>.

European Commission(2020b). White Paper on Data Ethics in Public Procurement of AI-based Services and Solutions. <https://ec.europa.eu/futurium/en/european-ai-alliance/white-paper-data-ethics-public-procurement-ai-based-services-and-solutions.html (접근일: 2022. 11. 1)>.

European Commission(2021). Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. COM(2021)206 final. <https://eur-lex.europa.eu/legal-content/EN/HIS/?uri=COM:2021:206:FIN (접근일: 2022. 11. 1)>.

Government of the Netherlands(2022). Impact Assessment Fundamental rights and algorithms. <https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms (접근일: 2022. 8. 15)>.

High-Level Expert Group on Artificial Intelligence(2019). Ethics guidelines for trustworthy AI. European Commission. <https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html (접근일: 2022. 8. 15.)>

High-Level Expert Group on Artificial Intelligence(2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. made public on the 17th of July 2020. <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (접근일: 2022. 8. 15)>.

Microsoft(2022). Responsible AI Impact Assessment Guide, June 2022. <https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Guide.pdf (접근일: 2022. 11. 1)>.

The Danish Institute for Human Rights(2020a). Human Rights Impact Assessment: Guidance and Toolbox. <https://www.humanrights.dk/tools/human-rights-impact-assessment-guidance-toolbox (접근일: 2022. 12. 1)>.

The Danish Institute for Human Rights(2020b). Guidance on Human Rights Impact Assessment of Digital Activities. <https://www.humanrights.dk/publications/human-rights-impact-assessment-digital-activities (접근일: 2022. 8. 15)>.

The Office of Science and Technology Policy(2022). Blueprint for an AI Bill of Rights : Making Automated Systems Work For The American People. The White House. <https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (접근일: 2022. 11. 1)>.

UNITED NATIONS(2011). Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework. UN Doc. A/HRC/17/31(21 March 2011). 번역본: 국가인권위원회(2011).

UNITED NATIONS(2018). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. UN doc. A/73/348(29 August 2018).

UNITED NATIONS(2021). The right to privacy in the digital age : Report of the United Nations High Commissioner for Human Rights. UN Doc. A/HRC/48/31(13 September 2021).

United Nations Human Rights Special Procedures(2020). How to Make Economic Reforms Consistent with Human Rights Obligations: Guiding Principles on Human Rights Impact Assessment of Economic Reforms. <https://www.ohchr.org/sites/default/files/GuidePrinciples_EN.pdf(접근일: 2022. 12. 1)>.

부 록

부록 I.

유럽연합 신뢰할 수 있는 인공지능 평가 목록[1)]

기본권

1. 인공지능 시스템은 잠재적으로 다음과 같은 근거 등에 기초하여 사람들을 부정적으로 차별합니까? (성별, 인종, 피부색, 민족적 또는 사회적 기원, 유전적 특징, 언어, 종교 또는 신념, 정치적 또는 다른 의견, 소수 민족의 일원, 재산, 출생, 장애, 나이 또는 성적지향 등)
   — 인공지능 시스템의 개발, 배치 및 사용단계에서 잠재적인 부정적인 차별(편향)을 테스트하고 모니터링하는 프로세스를 마련했습니까?
   — 인공지능 시스템에서 잠재적인 부정적인 차별(편향)을 해결하고 시정하기 위한 프로세스를 마련했습니까?

2. 인공지능 시스템은 예를 들어, 아동보호와 아동의 최선의 이익을 고려하는 것과 관련하여 아동의 권리를 존중합니까?
   — 인공지능 시스템이 아동에게 미칠 잠재적인 피해를 해결하고 시정하는 프로세스를 마련했습니까?
   — 인공지능 시스템의 개발, 배치 및 사용단계에서 아동에 대한 잠재적인 피해를 테스트하고 모니터링하는 프로세스를 마련했습니까?

3. 인공지능 시스템은 GDPR에 따라 개인과 관련된 개인정보를 보호합니까?

---

1) High-Level Expert Group on Artificial Intelligence(2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. made public on the 17th of July 2020. <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (접근일: 2022. 8. 15)>.

— 인공지능 시스템의 개발, 배치 및 사용단계와 관련하여, 그 목적과 관련된 처리작업의 필요성 및 비례성 평가를 포함하여, 개인정보보호 영향평가의 필요성을 구체적으로 평가할 수 있는 프로세스를 구축했습니까?

— 인공지능 시스템의 개발, 배치 및 사용단계와 관련하여, 안전장치, 보안 조치 및 개인정보보호를 보장하기 위한 메커니즘을 포함하여, 위험에 대처하기 위한 조치를 구축했습니까?

— 이 평가목록의 프라이버시 및 데이터 거버넌스 섹션과 유럽개인정보보호감독관(European Data Protection Supervisor)이 제공하는 지침을 참조하십시오.

4. 인공지능시스템은 표현과 정보의 자유 및 집회와 결사의 자유를 존중합니까?

— 인공지능 시스템의 개발, 배치 및 사용단계에서, 표현과 정보의 자유, 집회와 결사의 자유에 대한 잠재적인 침해를 테스트하고 모니터링하는 프로세스를 구축했습니까?

— 인공지능 시스템에서 표현과 정보의 자유 및 집회와 결사의 자유에 대한 잠재적인 침해를 해결하고 시정하기 위한 프로세스를 구축했습니까?

1. 인간행위자와 감독

○ 인간행위자와 자율성

- 인공지능 시스템은 최종사용자가 인간이나 사회에 영향을 미치는 상호작용, 안내 또는 결정을 내리도록 설계되었습니까?

— 인공지능 시스템이 결정, 내용, 조언 또는 결과가 알고리즘 결정의 결과인지 여부에 대해 일부 또는 모든 최종사용자 또는 주체에 혼란을 일으킬 수 있습니까?

— 최종사용자 또는 다른 주체들이 결정, 내용, 조언 또는 결과가 알고리즘 결정의 결과라는 것을 적절하게 인식하고 있습니까?

- 인공지능 시스템이 일부나 모든 최종사용자 또는 주체에게 인간 또는 인공지능 시스템과 상호작용하는지 여부에 대해 혼란을 일으킬 수 있습니까?
  — 최종사용자 또는 주체가 인공지능시스템과 상호작용하고 있다는 정보를 받습니까?
- 인공지능 시스템이 최종사용자의 과도한 의존도를 발생시킴으로써 인간의 자율성에 영향을 미칠 수 있습니까?
  — 최종사용자가 인공지능 시스템에 과도하게 의존하지 않도록 절차를 마련했습니까?
- 인공지능 시스템이 의도하지 않고 바람직하지 않은 방식으로 최종사용자의 의사결정과정에 간섭함으로써 인간의 자율성에 영향을 미칠 수 있습니까?
  — 인공지능 시스템이 실수로 인간의 자율성에 영향을 미치지 않도록 어떤 절차를 마련했습니까?
- 인공지능 시스템은 최종사용자 또는 주체간의 또는 그들과의 사회적 상호작용을 시뮬레이션합니까?
- 인공지능 시스템은 인간의 애착을 형성하거나, 중독적인 행동을 자극하거나, 사용자 행동을 조작할 위험이 있습니까? 어떤 위험이 가능한지 또는 가능성이 있는지에 따라 아래 질문에 답하십시오.
  — 인공지능 시스템에 불균형한 애착을 갖게 될 경우 최종사용자 또는 주체에게 발생할 수 있는 부정적인 결과에 대처하기 위한 조치를 취했습니까?
  — 중독의 위험을 최소화하기 위한 조치를 취했습니까?
  — 조작의 위험을 완화하기 위한 조치를 취했습니까?

○ 인간의 감독

- 인공지능 시스템이 다음과 같은지 여부를 확인하십시오(적절한 것을 모두 선택하세요).
  — 자가학습 또는 자율적인 시스템입니까?

— Human-in-the-Loop(인간참여)에 의해 감독됩니까?

— Human-on-the-Loop(인간지배)에 의해 감독됩니까?

— Human-in-Command(인간지휘)에 의해 감독됩니까?

- 인간(Human-in-the-Loop, Human-on-the-Loop, Human-in-Command)은 감독을 수행하는 방법에 대한 특정 훈련을 받았습니까?
- 최종사용자 또는 주체에 대한 인공지능 시스템의 바람직하지 않은 부작용에 대한 탐지 및 대응 메커니즘을 수립했습니까?
- 필요할 때 작업을 안전하게 중단하기 위한 '중지 버튼' 또는 관련 절차를 보장했습니까?
- 인공지능 시스템의 자가학습 또는 자율성을 반영하기 위해 특정 감독 및 통제 조치를 취했습니까?

2. 기술적 견고성과 안전성

○ 공격에 대한 회복성과 보안

- 인공지능 시스템이 설계 또는 기술적 결함, 결점, 중단, 공격, 오용, 부적절하거나 악의적인 사용과 같은 위험이나 위협의 경우에 (예를 들어, 인간 또는 사회 안전에) 적대적이고 치명적인 해로운 영향을 미칠 수 있습니까?
- 인공지능 시스템이 사이버보안에 대해 인증을 받았습니까? (예: 유럽의 사이버보안법에 의해 생성된 인증제도) 또는 특정 보안 표준을 준수합니까?
- 인공지능 시스템은 사이버공격에 얼마나 노출되어 있습니까?

  — 인공지능 시스템이 취약할 수 있는 잠재적인 공격 형태를 평가했습니까?

  — 다음과 같은 다양한 유형의 취약점과 잠재적인 공격 진입점을 고려했습니까?

  * 데이터 중독(즉, 훈련데이터 조작)

  * 모델 회피(즉, 공격자의 의지에 따라 데이터를 분류하는 것)

  * 모델 역전(즉, 모델 매개변수 추론)

- 수명주기 동안 발생할 수 있는 잠재적인 공격에 대비하여 인공지능 시스템의 무결성, 견고성 및 전반적인 보안을 보장하기 위한 조치를 취했습니까?
- 시스템을 레드팀(red-team)/펜테스트(침투시험,pentest)했습니까?
- 최종사용자에게 보안적용 범위 및 업데이트 기간을 알렸습니까?
  — 인공지능 시스템에 대한 보안 업데이트를 제공하는 예상 기간은 얼마나 됩니까?

○ 일반적 안전성

- 각 특정 용도 사례에서 인공지능 시스템의 위험, 위험지표 및 위험 수준을 정의했습니까?
  — 위험을 지속적으로 측정하고 평가하는 프로세스를 마련했습니까?
  — 최종사용자와 주체에게 기존 또는 잠재적 위험을 알렸습니까?
- 인공지능 시스템에 대한 가능한 위협(설계 결함, 기술적 결함, 환경 위협)과 발생할 수 있는 결과는 무엇입니까?
  — 인공지능 시스템의 악의적 사용, 오용 또는 부적절한 사용의 위험을 평가했습니까?
  — 인공지능 시스템의 결함 또는 오용으로 인해 발생할 수 있는 결과의 (예를 들어, 인간의 무결성과 관련된) 안전 임계 수준을 정의했습니까?
- 안정적이고 신뢰할 수 있는 동작에 대한 중요한 인공지능 시스템의 결정 의존도를 평가했습니까?
  — 신뢰성/테스트 요구사항을 적절한 수준의 안정성 및 신뢰성에 맞추었습니까?
- 중복시스템 또는 다른 병렬시스템(인공지능 기반 또는 '기존'시스템)을 통해 내결함성을 계획했습니까?
- 인공지능 시스템이 기술적 견고성과 안전성에 대한 새로운 검토를 받을 수 있도록 언제 변경되었는지 평가하는 메커니즘을 개발했습니까?

○ 정확성

- 인공지능 시스템의 정확도가 낮으면 치명적이거나 적대적이거나 해로운 결과를 초래할 수 있습니까?
- 인공지능 시스템을 개발하는데 사용되는 데이터(훈련데이터 포함)가 시스템이 배포될 환경을 대표하는 고품질의 완전하고 최신 상태인지 확인하기 위한 조치를 취했습니까?
- 인공지능 시스템의 정확성을 모니터링하고 문서화하기 위한 일련의 단계를 마련했습니까?
- 인공지능 시스템의 작동이 학습된 데이터나 가정을 무효화할 수 있는지 여부와 이것이 어떻게 적대적인 영향을 미칠 수 있는지 고려했습니까?
- 최종사용자 및 주체가 예상할 수 있는 인공지능 시스템의 정확도 수준이 적절하게 전달되도록 프로세스를 마련했습니까?

○ 신뢰성, 대체계획 및 재현성

- 신뢰도 및 재현성이 낮은 경우 인공지능 시스템이 (예를 들어, 인간의 안전과 관련된) 치명적, 적대적 또는 해로운 결과를 일으킬 수 있습니까?
  - 인공지능 시스템이 의도한 목표를 충족하는지 모니터링하기 위해 잘 규정된 프로세스를 마련했습니까?
  - 재현성을 보장하기 위해 특정 상황이나 조건을 고려해야 하는지 여부를 테스트했습니까?
- 인공지능 시스템의 신뢰성과 재현성의 다양한 측면을 평가하고 보장하기 위해 검증 및 승인 방법과 문서화(예:로깅)를 마련했습니까?
  - 인공지능 시스템의 신뢰성 및 재현성을 테스트하고 검증하기 위한 프로세스를 명확하게 문서화하고 운영가능하도록 했습니까?
- 어떤 출처에서든 인공지능 시스템 오류를 해결하기 위한 테스트 된 안전장치 대체 계획을 정의하고, 이를 촉발하기 위한 거버넌스 절차를 마련하였습니까?

- 인공지능 시스템이 낮은 신뢰도 점수로 결과를 산출하는 경우를 처리하기 위한 적절한 절차를 마련했습니까?
- 인공지능 시스템은 (온라인) 연속 학습(continual learning)을 사용하고 있습니까?
  — 인공지능 시스템이 객관적 기능에 대한 점수를 높이기 위해 새로운 방법이나 특이한 방법을 학습하는데 있어 잠재적인 부정적 결과를 고려했습니까?

3. 프라이버시 및 데이터 거버넌스

○ 프라이버시

- 프라이버시권, 신체적, 정신적 및 도덕적 무결성에 대한 권리, 개인정보보호 권리에 대한 인공지능 시스템의 영향을 고려했습니까?
- 사용 사례에 따라, 인공지능시 스템에 관한 프라이버시와 관련된 문제를 표시할 수 있는 메커니즘을 설정했습니까?

○ 데이터 거버넌스

- 인공지능 시스템이 개인정보(특정 범주의 개인정보 포함)를 사용하거나 처리하여 훈련을 받았거나 개발되었습니까?
- GDPR(일반개인정보보호법) 혹은 유럽 이외 지역에서 이에 상응하는 규정에 따라 일부는 의무적으로 적용되는, 다음 조치를 취했습니까?
  * 개인정보영향평가(DPIA)
  * 개인정보보호책임자(DPO)를 지정하고 그들을 인공지능 시스템의 개발, 조달 또는 사용 단계의 초기에 포함
  * 개인정보 처리를 위한 감독 메커니즘(자격 있는 직원으로 접근 제한, 데이터 접근 기록 및 수정 메커니즘 포함)
  * 설계 및 기본값에 의한 프라이버시 보호를 달성하기 위한 조치(예: 암호화,

가명화, 집계, 익명화)

* 데이터 최소화, 특히 개인정보(특정 범주의 개인정보 포함)

— 동의철회권, 거부권, 잊힐 권리를 인공지능 시스템 개발에 구현하였습니까?

— 인공지능 시스템의 수명주기 동안 수집, 생성 또는 처리되는 데이터의 프라이버시 및 개인정보보호 영향을 고려했습니까?

- 인공지능 시스템의 비개인 훈련데이터 또는 기타 처리된 비개인 데이터의 프라이버시 및 개인정보보호 영향을 고려했습니까?
- 인공지능 시스템을 관련 표준(예: ISO, IEEE) 또는 (일상)데이터 관리 및 거버넌스를 위해 널리 채택된 규약에 맞추었습니까?

4. 투명성

○ 추적가능성

- 전체 수명주기 동안 인공지능 시스템의 추적가능성을 다루는 조치를 취했습니까?
  — 인공지능 시스템에 대한 입력데이터의 품질을 지속적으로 평가하기 위한 조치를 마련했습니까?
  — 인공지능 시스템이 특정 결정(들) 또는 권장사항(들)을 내리는데 사용된 데이터를 추적할 수 있습니까?
  — 어떤 인공지능 모델이나 규칙이 인공지능 시스템의 결정(들)이나 권고사항으로 이어졌는지 추적할 수 있습니까?
  — 인공지능 시스템의 출력 품질을 지속적으로 평가하기 위한 조치를 취했습니까?
  — 인공지능 시스템의 결정(들) 또는 권장사항(들)을 기록하기 위해 적절한 로깅 관행을 마련했습니까?

○ 설명가능성

- 인공지능 시스템의 결정을 사용자에게 설명했습니까?
- 사용자가 인공지능 시스템의 결정을 이해하고 있는지 지속적으로 조사합니까?

○ 고지

- 대화형 인공지능 시스템(예: 챗봇, 로봇 변호사)의 경우, 사용자가 인간이 아닌 인공지능 시스템과 상호작용하고 있음을 사용자에게 전달합니까?
- 인공지능 시스템에 의해 생성된 결정의 목적, 기준 및 한계에 대해 사용자에게 알리는 메커니즘을 수립했습니까?
  - 인공지능 시스템의 이점을 사용자에게 전달했습니까?
  - 정확도 및 오류율과 같은 인공지능 시스템의 기술적 한계 및 잠재적 위험을 사용자에게 전달했습니까?
  - 인공지능 시스템을 적절하게 사용하는 방법에 대해 사용자에게 적절한 교육 자료 및 면책조항을 제공했습니까?

5. 다양성, 차별금지, 공정성

○ 불공정한 편향 회피

- 인공지능 시스템에서, 입력데이터 사용 및 알고리즘 설계 모두에 대해 불공정한 편향을 생성하거나 강화하는 것을 방지하기 위한 전략이나 일련의 절차를 수립했습니까?
- 데이터에서 최종사용자 및 주체의 다양성과 대표성을 고려했습니까?
  - 특정 대상그룹이나 문제가 있는 사용 사례에 대해 테스트했습니까?
  - 데이터와 모델 및 성능에 대한 이해를 향상시키기 위해, 공개적으로 사용가능한 최신 기술이 적용된 기술 도구를 조사하고 사용했습니까?

— 인공지능 시스템의 전체 수명주기 동안 잠재적 편향(예: 사용된 데이터셋 구성으로 발생할 수 있는 제한에 따른 편향(다양성 부족, 대표성 부족))에 대해 테스트 및 모니터링하는 프로세스를 평가하고 구축했습니까?

— 관련이 있는 경우, 데이터에서 최종사용자 및 주체의 다양성과 대표성을 고려했습니까?

- 인공지능 설계자와 인공지능 개발자가 인공지능 시스템을 설계하고 개발할 때 주입할 수 있는 편견을 더 잘 인식할 수 있도록 교육 및 인식 이니셔티브를 마련했습니까?
- 인공지능 시스템의 편향, 차별 또는 성능저하와 관련된 문제를 표시(신고)할 수 있는 메커니즘을 보장했습니까?

— 그러한 문제를 누구에게 어떻게 제기할 수 있는지에 대해 명확한 단계와 방법을 설정했습니까?

— (최종)사용자 및 주체 외에 인공지능 시스템에 의해 잠재적으로 직(간)접적으로 영향을 받을 수 있는 대상을 식별했습니까?

- 공정성에 대한 정의는 인공지능 시스템을 설정하는 프로세스의 모든 단계에서 일반적으로 사용되고 구현됩니까?

— 이것을 선택하기 전에 공정성에 대한 다른 정의를 고려했습니까?

— 공정성에 대한 올바른 정의에 대해 영향을 받는 지역 공동체, 즉 노인 또는 장애인 대표 등과 협의했습니까?

— 적용된 공정성의 정의를 측정하고 테스트하기 위한 정량적 분석 또는 측정기준을 확인했습니까?

— 인공지능 시스템의 공정성을 보장하기 위한 메커니즘을 구축했습니까?

○ 접근가능성 및 보편적 설계

- 인공지능 시스템이 사회의 다양한 선호도와 능력에 부합하는지 확인했습니까?
- 인공지능 시스템의 사용자 인터페이스가 특별한 필요나 장애가 있는 사람들, 또는

배제 위험이 있는 사람들이 사용할 수 있는지 평가했습니까?

— 인공지능 시스템에 대한 정보와 인공지능 시스템의 사용자 인터페이스가 보조기술(예:화면 판독기) 사용자에게도 접근하고 사용할 수 있도록 보장했습니까?

— 인공지능 시스템의 계획 및 개발단계에서 보조기술이 필요한 최종사용자 또는 주체를 참여시키거나 협의했습니까?

- 해당되는 경우, 계획 및 개발 과정의 모든 단계에서 보편적 설계 원칙을 고려했는지 확인했습니까?
- 인공지능 시스템이 잠재적인 최종사용자 및 주체에 미치는 영향을 고려했습니까?

— 인공지능 시스템 구축에 관련된 팀이 가능한 대상 최종사용자 및 주체와 협력했는지 평가했습니까?

— 인공지능 시스템의 결과에 불균형적으로 영향을 받을 수 있는 그룹이 있을 수 있는지 평가했습니까?

— 최종사용자 또는 주체의 공동체에 대한 시스템의 불공정성의 위험을 평가했습니까?

○ 이해관계자 참여

- 인공지능 시스템의 설계 및 개발에 가능한 가장 광범위한 이해관계자의 참여를 포함하는 메커니즘을 고려했습니까?

6. 사회·환경적 복지

○ 환경적 복지

- 인공지능 시스템이 환경에 미치는 잠재적인 부정적인 영향이 있습니까?

— 어떤 잠재적 영향을 파악합니까?

- 가능한 경우 인공지능 시스템의 개발, 배포 및 사용이 환경에 미치는 영향(예를 들어, 사용된 에너지양 및 탄소배출량)을 평가하는 메커니즘을 설정했습니까?
  — 인공지능 시스템의 수명주기 전반에 걸쳐 환경영향을 줄이기 위한 조치를 정의했습니까?

○ 일자리 및 역량에 대한 영향

- 인공지능 시스템이 인간의 업무와 업무배치에 영향을 미칩니까?
- 영향을 받는 노동자와 그 대표자(노동조합,(유럽)노동위원회)에 사전에 알리고 협의함으로써 조직 내 인공지능 시스템 도입을 위한 기반을 마련했습니까?
- 인공지능 시스템이 인간의 업무에 미치는 영향을 잘 이해할 수 있도록 조치를 취했습니까?
  — 노동자가 인공지능 시스템이 어떻게 작동하는지, 어떤 기능이 있고 어떤 기능이 없는지에 대해 이해하고 있는지 확인했습니까?
- 인공지능 시스템이 노동력의 탈숙련화 위험을 발생시킬 수 있습니까?
  — 탈숙련화 위험에 대응하기 위한 조치를 취했습니까?
- 시스템이 새로운 (디지털) 기술을 촉진하거나 필요로 합니까?
  — 기술 재교육과 향상을 위한 훈련 기회와 자료를 제공했습니까?

○ 사회 전반 및 민주주의에 대한 영향

- 인공지능 시스템이 사회 전반 및 민주주의에 부정적인 영향을 미칠 수 있습니까?
  — 잠재적으로 간접적으로 영향을 받는 이해관계자 또는 사회 전체와 같이 (최종)사용자와 주체를 넘어서는 인공지능 시스템 사용이 미치는 사회적 영향을 평가했습니까?
  — 인공지능 시스템의 잠재적인 사회적 피해를 최소화하기 위한 조치를 취했습니까?

— 인공지능 시스템이 민주주의에 부정적인 영향을 미치지 않도록 조치를 취했습니까?

7. 책무성

○ 감사가능성

- 인공지능 시스템의 감사가능성을 용이하게 하는 메커니즘(예를 들어, 개발 프로세스 추적가능성, 훈련데이터의 소싱 및 인공지능 시스템 프로세스, 결과, 긍정적/부정적 영향의 로깅)을 설정했습니까?
- 독립적인 제3자가 인공지능 시스템을 감사할 수 있는지 확인했습니까?

○ 위험관리

- 윤리적 문제 및 책무성 조치를 감독하기 위한 외부지침 또는 제3자 감사 프로세스를 미리 확인했습니까?
  — 이러한 제3자의 참여가 개발단계를 넘어선 것입니까?
- 위험 훈련을 조직했습니까? 그렇다면 인공지능 시스템에 적용가능한 잠재적인 법적 프레임워크에 대한 정보도 제공합니까?
- 잠재적으로 불분명한 영역을 포함하여, 전반적인 책무성 및 윤리 관행을 논의하기 위해 인공지능 윤리 검토위원회 또는 유사한 메커니즘을 설립하는 것을 고려했습니까?
- 이 ALTAI(신뢰할 수 있는 인공지능 평가목록)에 대한 인공지능 시스템의 준수 여부를 논의하고 지속적으로 모니터링 및 평가하는 프로세스를 수립했습니까?
  — 이 프로세스에 앞서 언급한 6가지 요구사항 간의 충돌 또는 서로 다른 윤리 원칙 간의 충돌에 대한 식별 및 문서화, 그리고 ‘절충(trade-off)’ 결정에 대한 설명이 포함됩니까?

— 그러한 프로세스에 관련된 사람들에게 적절한 훈련을 제공했으며 이는 인공지능 시스템에 적용할 수 있는 법적 프레임워크도 포함합니까?

- 제3자(예: 공급업체, 최종사용자, 주체, 유통업체/판매업체 또는 노동자)가 인공지능 시스템의 잠재적인 취약성, 위험 또는 편향을 보고하는 프로세스를 수립했습니까?

  — 이 프로세스가 위험관리 프로세스의 개정을 촉진합니까?

- 개인에게 부정적인 영향을 미칠 수 있는 응용프로그램에 대해 설계에 의한 시정 메커니즘(redress by deisng mechanisms)이 마련되어 있습니까?

부록 Ⅱ.

미국 인공지능 권리장전 청사진[2)]

안전하고 효과적인 시스템

우리는 안전하지 않거나 효과적이지 못한 시스템으로부터 보호되어야 합니다.

자동화된 시스템은 시스템의 우려사항, 위험성 및 잠재적 영향을 식별하기 위해 다양한 집단, 이해관계자 및 분야별 전문가의 자문을 받아 개발되어야 합니다. 시스템에 대하여 사전적인 배치 테스트, 위험 식별 및 완화 조치, 지속적인 모니터링을 수행하여, 시스템이 의도된 대로 사용되어 안전하고 효과적이며, 의도된 사용을 벗어나는 안전하지 않은 결과물에 대해서는 완화조치를 취하였고, 분야별 표준을 준수하였다는 사실을 입증하여야 합니다. 이러한 보호 조치로 시스템을 도입하지 않거나 시스템 사용을 중단할 수도 있어야 합니다. 자동화된 시스템이 여러분의 안전이나 지역사회의 안전을 위협하려는 의도로, 또는 그러할 가능성이 합리적으로 예측되는 방식으로 설계되어서는 안 됩니다. 자동화된 시스템은 의도되지 않았으나 예측가능했던 사용 또는 영향으로부터 여러분을 사전적으로 보호할 수 있도록 설계되어야 합니다. 여러분은 자동화된 시스템의 설계, 개발 및 배치 시 부적절하거나 관련 없는 데이터의 사용으로부터 보호받아야 하며, 그 재사용으로 인한 복합적인 위해로부터도 보호받아야 합니다. 시스템이 안전하고 효과적이라는 사실을 확인하는 독립적인 평가와 보고가 수행되어야 하며, 여기에는 잠재적 위해를 완화하기 위해 취한 조치에 대한 보고가 포함되어야 하고, 그 결과는 가능할 한 공개되어야 합니다.

알고리즘 차별로부터 보호

2) The Office of Science and Technology Policy(2022). Blueprint for an AI Bill of Rights : Making Automated Systems Work For The American People. The White House. <https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (접근일: 2022. 11. 1)>.

여러분은 알고리즘에 의해 차별받지 않아야 하며, 시스템은 공평한 방식으로 사용되고 설계되어야 합니다.

알고리즘 차별이란, 자동화된 시스템이 인종, 피부색, 민족, 성별(임신, 출산 및 관련 건강 상태, 성 정체성, 간성 상태, 성적 지향 포함), 종교, 연령, 출신 국가, 장애, 퇴역 상태, 유전 정보, 기타 법률이 민감정보로 분류하여 보호하는 특성에 기반하여 개인에 대해 비합리적으로 다르게 대우하거나 불리한 영향을 줄 때 발생합니다. 특정 상황에서는 이러한 알고리즘 차별이 위법할 수 있습니다. 자동화된 시스템을 설계, 개발 및 도입하는 기관은 알고리즘 차별로부터 개인과 지역사회를 보호하고 시스템을 공평한 방식으로 사용하고 설계할 수 있도록 사전적이고 지속적인 조치를 취해야 합니다. 이러한 보호 조치로서, 시스템 설계, 대표 데이터 사용 및 인구집단 특성별 대리변수 보호에 대한 사전 공평성 평가를 수행하고, 설계 및 개발에 대한 장애인에 대한 접근성을 보장하며, 배치 전 및 지속적으로 편향성 테스트 및 완화조치를 수행하고, 조직적 감독을 명확하게 수행하여야 합니다. 알고리즘영향평가 형식의 독립적인 점검과 쉬운 용어로 이루어진 보고가 이루어져야 하며, 여기에는 편향성 테스트 결과 및 완화 조치에 대한 정보가 포함되어야 하고, 이러한 보호 조치 여부를 확인하기 위해 보고서가 가능한 한 공개되어야 합니다.

개인정보 보호

여러분은 설치형 보호(Built-In Protection)를 통해 개인정보를 오남용하는 관행으로부터 보호받아야 하고, 여러분에 관한 개인정보가 어떻게 이용되는지에 대해 여러분이 권한을 가져야 합니다.

여러분은 프라이버시 보호장치가 기본설정으로 보장되는 설계를 선택함으로써 프라이버시 침해로부터 보호받아야 하며, 이러한 설계에는 개인정보 수집에 있어 특정 상황에 꼭 필요한 개인정보만 수집될 것이라는 합리적인 기대에 부응하는 설계가 포함됩니다. 자동화된 시스템을 설계, 개발 및 도입하는 기관은 여러분의 개인정보에 대한 수집, 이용, 접근, 전송 및 삭제에 대해 적절한 방식으로 최대한 여러분의 허락을 구해야 하며, 여러분의 결정을 존중해야 하며, 이것이 가능하지 않은 경우 대안적인 프라이버시 기본 설계 보호장치를 사용해야 합니다. 시스템은 이용자 선택을 모호하게 하거나 개인정보를

침해하는 기본설정으로 이용자에게 부담을 지우는 이용자 경험 및 설계상 결정을 사용해서는 안 됩니다. 동의는 적절하고 유의미하게 이루어질 수 있는 경우에만 개인정보수집을 정당화하는 근거로 사용되어야 합니다. 모든 동의 요청은 간단하고 이해하기 쉬운 용어로 작성되어야 하며 개인정보가 수집되고 특정 상황에서 사용되는 데 대한 결정 권한이 여러분에게 주어져야 합니다. 개인정보의 광범위한 이용을 요구하면서 현재처럼 이해하기 어려운 방식으로 통지하고 선택하도록 하는 관행은 변화해야 합니다. 건강, 직업, 교육, 형사 사법, 금융을 포함하는 민감한 영역과 관련된 개인정보처리 및 추론, 그리고 아동청소년 관련 개인정보에 대한 보호 강화 및 제한을 최우선으로 고려하여야 합니다. 민감한 영역에서 여러분의 개인정보 및 관련된 추론은 필수적인 기능 용도로만 사용되어야 하며 윤리적 검토 및 사용 금지 조치로 보호받을 수 있어야 합니다. 여러분과 여러분이 속한 집단은 견제되지 않은 감시로부터 자유로울 수 있어야 합니다. 감시 기술이 사생활과 시민권에 미치는 잠재적 위해와 이를 제한하는 범위를 최소한 배치 전에 평가하는 등 감시 기술에 대한 감독을 강화하여야 합니다. 지속적인 감시 및 모니터링 기술이 교육, 직업, 주택 분야에서 사용되거나, 그러한 감시 기술의 사용이 개인의 권리, 기회 또는 접근을 제한할 가능성이 있는 여타의 상황에서 사용되어서는 안 됩니다. 여러분은 가능한 한 여러분의 개인정보 결정권이 존중되었음을 확인하고 여러분의 권리, 기회 또는 접근에 미치는 감시 기술의 잠재적 영향에 대해 평가하는 보고서에 접근할 수 있어야 합니다.

통지 및 설명

여러분은 자동화된 시스템이 사용되고 있다는 사실과 이 시스템이 여러분에게 영향을 미치는 결과물에 작동하는 방법과 이유를 이해할 수 있어야 합니다.

자동화된 시스템을 설계, 개발 및 도입하는 기관은 전반적인 시스템 기능 및 자동화가 수행하는 역할에 대한 명확한 설명, 시스템이 사용 중이라는 통지, 시스템을 책임지는 개인이나 기관, 그 결과물에 대하여 명확하고 시기적절하며 공개적인 설명 등을 쉬운 용어로 서술한 문서를 일반에 공개하여야 합니다.

이러한 통지는 최신 상태로 유지되어야 하며, 시스템의 영향을 받는 사람들은 중요한

사용 사례 또는 주요 기능 변경에 대한 통지를 받아야 합니다. 여러분은 자동화된 시스템이 여러분에게 영향을 미치는 결과물을 결정한 방법과 이유를 알 수 있어야 하며, 이때 자동화된 시스템이 결과물을 결정하는 유일한 입력이 아닌 경우에도 이 권리를 보장받을 수 있어야 합니다. 자동화된 시스템은 여러분과 시스템을 이해해야 하는 운영자 및 기타 사용자에게 기술적으로 유효하고 유의미하며 유용한 설명을 제공해야 하며, 상황에 따라 위험 수준이 조정되어야 합니다. 이러한 자동화된 시스템에 대한 요약 정보가 쉬운 용어로 포함된 보고서와 더불어, 이들 통지 및 설명의 명확성과 품질에 대한 평가 내용이 가능한 한 공개되어야 합니다.

인간 대안, 검토 및 대체

적절한 경우 여러분은 제외(opt-out)될 수 있어야 하며 여러분이 직면한 문제를 신속하게 검토하고 구제할 수 있는 사람에게 연락할 수 있어야 합니다.

적절한 경우 여러분은 인간 대안을 선택하고 자동화된 시스템에서 제외될 수 있어야 합니다. 이 적절성은 주어진 상황에 대한 합리적인 기대에 기반하고 광범위한 접근성을 보장하며 특히 위해적 영향으로부터 일반 사람들을 보호하는 데 중점을 두고 결정되어야 합니다. 경우에 따라 법률이 인간 대안 또는 다른 대안을 요구할 수 있습니다. 자동화된 시스템이 실패하거나 오류가 발생하였을 때, 혹은 여러분이 여러분에게 미친 영향에 대해 이의를 제기하거나 불복하고자 할 때, 여러분은 대체 절차 및 재심 절차를 통해 시기적절한 인간 검토 및 구제수단에 접근할 수 있어야 합니다. 인간 검토 및 대체는 접근 가능하고, 공평하고, 효과적이며, 계속 유지되고, 적절한 직원 교육이 수반되어야 하며, 일반 사람들에게 불합리한 부담을 주어서는 안 됩니다. 형사 사법, 고용, 교육 및 보건의료 또는 그 밖의 민감한 영역에서 의도적으로 사용되는 자동화된 시스템은 해당 목적에 추가적으로 맞춤하여야 하고, 감독기구에 유의미한 접근을 보장하여야 하며, 시스템과 상호 작용하는 모든 사람에 대한 교육을 실시하고, 부정적인 결정이나 고위험 결정에 대한 인간의 검토를 포함하여야 합니다. 이러한 인간 거버넌스 절차에 대한 설명과 적시성, 접근성, 결과물 및 효과성에 대한 평가 보고서는 가능한 한 공개되어야 합니다. □

부록 Ⅲ.

영국 NMIP 알고리즘영향평가 템플릿[3)]

프로젝트 이름: [삽입]

알고리즘영향평가 시작 날짜: [삽입], 마지막 편집: [삽입]

조직/팀 이름: [삽입]

프로젝트 연락처: [삽입]

접근성 및 사용성에 대한 참고사항

템플릿 버전: 1.0

이것은 제안된 NMIP 알고리즘영향평가 템플릿의 첫 번째 초안(버전 1.0)입니다. 우리는 접근성과 사용성을 목표로 했지만 이 프레임워크가 실제로 구현되면 다른 요구사항이 발생할 수 있습니다. 따라서 사용자 경험을 개선할 수 있는 디자인 또는 접근성 기능과 관련된 모든 피드백을 환영합니다: hello@adalovelaceinstitute.org

이 템플릿의 복사본을 다운로드하려면 파일 - 복사본 만들기로 이동합니다.

---

3) Ada Lovelace Institute(2020). Algorithmic impact assessment: AIA template. <https://www.adalovelaceinstitute.org/resource/aia-template/ (접근일: 2022. 8. 15)>.

목차

개요

이 문서는 NHS의 국가의료이미지플랫폼(National Medical Imaging Platform, NMIP)을 위한 알고리즘영향평가를 완료하기 위한 템플릿입니다

이 문서에서 NMIP에 액세스하려는 프로젝트 팀은 알고리즘영향평가를 구성하는 세 가지 수행의 결과를 보고합니다.

1. 초기 성찰적 수행(reflexive exercise), 프로젝트 팀이 완료하며, 프로젝트의 잠재적 영향에 대한 것입니다.
2. 참여형 워크숍, 프로젝트 팀이 환자 및 임상의 패널로부터 이 문서에 대한 피드백을 받습니다.
3. 성찰적 수행 및 참여형 워크숍의 최종 종합, 이는 NMIP 데이터 액세스 위원회(DAC)에 제출되며, NMIP에 대한 액세스를 승인하거나 거부하는 최종 결정의 일부로 고려됩니다.

이 문서는 이 템플릿을 사용하는 방법과 프로젝트 팀이 수행해야 하는 알고리즘영향평가 프로세스의 기타 활동에 대한 자세한 설명을 제공하는 NMIP 알고리즘영향평가 사용자 가이드와 함께 읽어야 합니다.

목적

이 템플릿에는 다음을 위한 일련의 질문과 답변 프롬프트가 포함되어 있습니다.

- 당신이 프로젝트에 대한 상위 수준 정보를 설명하는 데 도움
- 의료 AI 프로젝트가 직면할 수 있는 몇 가지 일반적인 윤리적 문제 고려
- 잠재적 위험의 식별 및 완화를 위해 사용 중인 귀하의 시스템의 여러 가능한 시나리오 구성

이러한 프롬프트는 팀이 영향을 식별하고 평가하는 데 도움이 되는 장치 역할을 합니다. 이것의 목적은 NMIP DAC에 신청 절차의 일부로 검토할 수 있는 영향 식별의 구체적인 기록을 제공할 뿐만 아니라, 당신의 팀이 시스템을 보다 효과적이고 안전하며 성공적으로 만드는 방법을 식별하는 데 도움이 되는 것입니다.

3단계를 모두 거친 알고리즘영향평가는 투명성과 신뢰성에 대한 약속의 일환으로 NHS AI Lab 웹사이트에 게시됩니다.

기대

이 템플릿의 프롬프트는 높은 수준의 프로젝트 정보 섹션부터 시작하여 시간순으로 작업하도록 설계되었습니다.

일부 질문에는 성찰적 수행, 참여적 워크숍 및 종합 섹션이라는 제목 아래에 답변을 위한 3개의 글머리 기호가 있습니다. 이는 알고리즘영향평가 절차의 개별 단계를 반영하고, 프로젝트의 다음 단계에서 답변으로 돌아가야 하는 위치를 구분합니다(위의 프로세스 그림 참조). 예를 들어, 성찰적 수행에 대한 답변은 '성찰적 수행'으로 표시된 행에 채워야 하며, 참여적 워크숍의 새로운 관찰은 '참여적 워크숍' 행에 통합되어야 합니다. 두 실행이 모두 완료된 후의 최종 평가 및 종합은 최종 '종합' 행에 문서화됩니다.

참고:

- 템플릿의 모든 필드는 필수 입력 사항입니다.
- 가능하면 모든 답변을 250단어 미만으로 유지하십시오. 질문은 예상되는 세부 수준에 대한 추가 지침을 제공합니다.
- 팀은 이 템플릿을 작성할 때 전문 용어, 임상 및/또는 기술 용어를 피하고(이러한 용어가 필요한 경우 명확하게 설명해야 함) 사전 이해를 전제로 하지 않고 평이한 일반 언어를 사용해야 합니다. 길에서 만난 낯선 사람에게 답을 설명한다고 상상하면서 답의 명확성을 스트레스 테스트할 수 있습니다. 기술 프로젝트의 일반 언어 요약에 대한 추가 팁은 일반 언어 요약에 대한 이 지침을 보세요.

1. 상위 수준의 프로젝트 정보

이 섹션에서는 당신의 프로젝트에 익숙하지 않은 이 알고리즘영향평가 검토자를 위해 프로젝트에 대한 배경 정보를 묻습니다. 또한 이 템플릿의 후반 섹션에서 윤리적 고려 사항 및 잠재적인 피해에 대한 생각에 도움을 주는 상위 수준의 맥락 질문을 다룹니다.

귀하의 프로젝트

1.ai) 귀하의 프로젝트 목적을 설명하십시오. 이것은 250단어 이내의 간결한 요약이어야 합니다. 이것을 논문 초록의 형태로 쓸 수 있습니다. 독자가 기술 지식이 많지 않다고 가정하십시오. 당신은 이것을 낯선 사람에게 설명하고 있습니다.

1.a.ii) 프로젝트의 의도된 용도를 설명합니다.

귀하의 조직

1.bi) 귀하의 조직에 대한 한 줄의 설명을 제공하십시오. 웹사이트 링크가 있는 경우 여기에 포함할 수도 있습니다.

1.b.ii) 어떤 유형의 조직/프로젝트 팀입니까? 예: 학술 연구실, 비영리 단체, 회사.

1.b.ii) 조직이나 팀에 사명 선언문이 있는 경우 여기에 포함할 수 있습니다.

1.c 이 시스템, 모델 또는 연구에 대한 입력 및 출력을 설명합니다. 어떤 다른 종류의 데이터 소스를 사용할 것입니까?

1.d 이 시스템의 영향을 받는 이해 관계자는 누구입니까? 예를 들어, 의도된 사용자는 누구이며, 누구에게 서비스를 제공하는 것입니까? 이를 나열할 때 가능한 한 구체적으로 작성하세요. 예를 들어 임상의, 간호사, 병원 행정 직원, 특정 유형의 환자 등 입니까? 영향을 받는 인구를 명시적으로 설명하고, 필요한 경우 시스템 사용자 집합과 구별해야 합니다.

성찰적 수행	참여 워크숍	종합

2. 통상적인 윤리적 고려 사항

이 섹션에서는 의료, AI 및 알고리즘 문헌의 맥락에서 일반적인, 특정 윤리적 고려 사항을 안내합니다.

이 섹션 작성에 도움이 필요한 경우, NHSX의 AI 보고서 섹션 3을 참조하여 의료 분야에서 AI를 사용하기 위한 윤리적 고려 사항뿐만 아니라, 보다 구체적으로는 데이터/디지털 건강 및 알고리즘 고려 사항을 참조할 수 있습니다. 이는 직간접적인 이해 관계자에 대한 당신의 모델의 가능한 영향을 분석하는 데 도움이 될 것입니다. 예를 들어, 프롬프트 2b에서, 당신은 사생활과 데이터 공유, 그리고 환자 안전에 대한 가능한 영향에 대해 논의할 수 있습니다.

2.a 이 프로젝트가 특정 커뮤니티에 대한 불평등 또는 불법적인 차별의 생성 또는 악화로 이어질 수 있습니까? 예를 들어, 치료에 대한 차별적 접근을 악화시키면서? 편향 및 공정성을 평가하거나 모니터링하기 위한 현재 계획에서 간과할 수 있는 것은 무엇입니까?

성찰적 수행	참여 워크숍	종합

2.b 귀하의 프로젝트는 동의와 자율성을 어떻게 고려합니까? 감시 증가와 관련된 위험이 있습니까? 예를 들어, 시스템의 의도된 수혜자에게 시스템 사용에 대해 어떻게 알립니까? 이 시스템은 감시가 증가하는 것으로 해석될 수 있습니까?

성찰적 수행	참여 워크숍	종합

2.d 이 프로젝트는 어떤 환경적 영향을 미칠까요? 이 시스템을 학습시키고 실행하는 데 얼마나 많은 컴퓨팅과 에너지가 필요합니까? 이 시스템의 결과로 발생할 수 있는 소프트웨어, 하드웨어 또는 장비의 다른 환경적 영향이 있습니까?

성찰적 수행	참여 워크숍	종합

2.e 이 시스템의 사용이 환자와 건강 및 치료 전문가 간의 관계에 어떤 영향을 미칠 수 있습니까? 이 시스템으로 인해 일부 환자 또는 서비스 사용자가 치료를 덜 받거나 건강 및 치료 전문가와 논의할 때 덜 솔직해질 수 있습니까?

성찰적 수행	참여 워크숍	종합

2.f 이 시스템을 사용하거나 영향을 받는 개인은 결과에 대해 어떻게 이의를 제기할 수 있습니까? 결과에 어떻게 이의를 제기할 수 있습니까? 이 시스템 사용을 거부할 수 있는 옵션이 있습니까? 이 시스템의 결과물은 어떻게 그리고 누구에게 설명 및 해석 가능합니까?

성찰적 수행	참여 워크숍	종합

2.e 어떻게 이 시스템이 의도하지 않거나 의도적으로 오용될 수 있습니까? 이 시스템이 사고나 오류로 이어질 수 있는 경우는 무엇입니까? 건강 및 사회복지 데이터 사용의 결과가 공익에 부합하지 않는 목적으로 다뤄질 수 있습니까?

성찰적 수행	참여 워크숍	종합

3. 영향 식별 및 시나리오

이 섹션은 구현 시 시스템의 광범위한 잠재적 영향을 반영하기 위한 것입니다.

위의 질문을 염두에 두고, 이 시스템이 보급된 후 이 시스템의 사용으로 인해 발생할 수 있는 최상의 시나리오와 최악의 시나리오, 성공을 위해 필요한 사회 환경 조건 및 극복해야 할 예상되는 장애물이나 과제는 무엇인지를 쉬운 언어로 요약하여 제공하세요.

여기에 기술 개념을 사용해야 하는 경우 주의 깊게 설명해야 합니다.

- 영향을 받는 이해 관계자에 대해 생각할 때, 직접적인 이해 관계자(예: 임상의, 환자, 의도된 환경에서 기술의 다른 사용자)와 간접 이해 관계자(예: 특정 정체성 그룹, 규제 기관, 시민 사회, 일반 대중)를 모두 고려하십시오.
- 이 프로세스에서 식별되어야 하는 영향의 최소 개수는 없지만, DAC는 최상의 시나리오와 최악의 시나리오 모두에서의 분석을 보기를 기대하고 있습니다. 위원회는 당신의 프로젝트에 대한 지식과 참여 실행 워크숍에 대한 참여의 강도에 기반하여 당신 팀의 평가에 동의할지 여부를 판단할 것입니다.

3.a 이 시스템을 사용할 때 발생할 수 있는 최상의 시나리오는 무엇입니까? 이 시스템이 설계/의도된 대로 작동할때 뿐만 아니라 실패, 오류, 실수 또는 예기치 않은 동작을 처리하는 방법에 대해서도 논의해야 합니다

성찰적 수행	참여 워크숍	종합

3.b 이 시스템이 성공적으로 운영되기 위해서는 어떤 사회-환경적 요구사항이 필요합니까? 예를 들어 인터넷에 대한 안정적인 연결, 의사와 간호사를 위한 교육, 특정 임상 및 행정 직원 간의 협업 등

이 질문에 답할 때, 어떤 이해 관계자가 이 시스템을 사용할 것인지, 시스템이 성공하기 위해 그들이 어떻게 최적으로 상호 작용하거나 협력할 것인지, 정보가 어떻게 (그리고 누구와) 공유되는지, 어떤 사회적, 기술적 및 작업흐름 의존성이 존재할 필요가 있는지 고려하세요. 또한 이 시스템을 성공적으로 사용하는 데 필요한 인프라 이해 관계자의 유형을 고려할 수도 있습니다.

성찰적 수행	참여 워크숍	종합

3.c 최상의 시나리오를 달성하기 위한 도전/장애물은 무엇입니까?

성찰적 수행	참여 워크숍	종합

3.d 이 시스템을 사용할 때 발생하는 최악의 시나리오는 무엇입니까?

3.di) 시스템이 설계/의도된 대로 작동할 때

성찰적 수행	참여 워크숍	종합

3.d.ii) 시스템이 설계/의도한 대로 작동하지 않거나 시스템이 작동하지 않을 때

성찰적 수행	참여 워크숍	종합

4. 잠재적 피해 분석

이 섹션은 프로젝트 팀에게 상기 3단계에서 설명한 모든 시나리오에서 발생하는 다양한 이해관계자의 잠재적 피해와 편익을 나열하도록 요청합니다.

그런 다음 팀이 각 피해의 인지된 중요성, 긴급성, 어려움 및 탐지 가능성이 무엇이라고 생각하는지 평가하도록 요청합니다.

마지막으로, 이 섹션에서는 프로젝트 팀이 이러한 영향이 피해를 유발할 가능성을 줄이는 특정 설계 결정과 같이, 이러한 피해에 대한 잠재적 완화를 고려하도록 요청합니다. 이것은 팀이 피해가 어떻게 분산될 수 있는지 고려하고, 잠재적인 즉각성 또는 복구 불가능성을 기반으로 어떤 피해에 우선 순위를 지정해야 하는지 확인하는 데 도움이 됩니다.

4.a 위에서 식별한 시나리오를 기반으로, 팀이 적극적으로 설계할 때 고려해야 하는, 이 시스템의 구현으로 인해 발생할 수 있는 잠재적인 피해는 무엇입니까? 누가, 어떻게 피해를 입을 위험이 가장 높습니까?

피해에 대해 생각할 때 임상의, 환자 또는 이 기술의 사용에 영향을 받을 다른 사람의 관점에서 이를 고려하십시오.

식별된 각각의 피해에 대해 다음 고려 사항을 기록해 두십시오.

- 중요도 - 이 피해가 이해관계자의 복지에 얼마나 결정적입니까? 회복할 수 없고 심각한 것은 무엇입니까?
- 긴급성 - 이 위협은 얼마나 즉각적입니까?
- 어려움 - 이 피해를 완화하는 것이 얼마나 어려울 것입니까?
- 탐지 가능성 - 현재 설계에서 이 피해를 얼마나 인지할 수 있습니까?

성찰적 수행	참여 워크숍	종합

4.b 위에서 확인된 시나리오를 기반으로, 이러한 피해를 최소화하기 위해 어떤 완화 조치를 취할 수 있습니까?

성찰적 수행	참여 워크숍	종합

부록 Ⅳ.

네덜란드 기본권 알고리즘영향평가 질의 문항[4)]

단계	항목 주제	질의
1부: 왜 하는가?	1.1 이유 및 문제 정의	1.1.1 알고리즘 사용/도입 계획에 대하여 설명하십시오. 알고리즘을 어떤 문제에 대한 해결책으로 예정하고 있습니까? 알고리즘을 사용하게 된 실제 사례 또는 이유는 무엇입니까? 왜 이 문제가 알고리즘을 필요로 합니까?
	1.2 목적	1.2.1 알고리즘을 사용하여 달성해야 하는 목적은 무엇입니까? 여기서 주 목적은 무엇이며 부차적 목적은 무엇입니까?
	1.3 공공 가치	1.3.1 알고리즘 사용이 촉발시킨 공공 가치는 무엇입니까? 알고리즘 사용이 촉발시킨 공공 가치가 다수일 때, 우선순위를 매길 수 있습니까? 1.3.2 알고리즘 사용의 결과로 손상을 입을 수 있는 공공 가치가 있습니까?
	1.4 법적 근거	1.4.1 알고리즘 사용의 법적 근거와 알고리즘에 기반해서 내려질 결정의 법적 근거는 무엇입니까?
	1.5 이해관계자 및 책임성	1.5.1 알고리즘 개발/사용/유지보수에 어떤 부서와 사람이 참여하고 있습니까? 1.5.2 알고리즘 개발 및 사용에 대한 책임이 투명하게 할당되었습니까? 알고리즘 개발이 완료되고 사용되는 경우 이러한 책임이 계속 투명하게 할당되도록 어떻게 보장될 수 있습니까? 1.5.3 알고리즘에 대한 궁극적인 책임이 누구에게 있습니까?

4) Government of the Netherlands(2022). Impact Assessment Fundamental rights and algorithms. <https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms (접근일: 2022. 8. 15)>.

2부: 무엇을 하는가?	2A (데이터 - 입력)	2A.1 평가: 알고리즘 유형	2A.1.1 어떤 유형의 알고리즘이 사용될 것인지 이미 (대략적으로) 알고 있습니까?
		2A.2 데이터 원천 및 품질	2A.2.1 알고리즘에 대한 입력으로 어떤 유형의 데이터가 사용되며 해당 데이터는 어떤 소스에서 가져 왔습니까? 입력 데이터가 사용되지 않으면 주제 2A.4로 진행하십시오. 2A.2.2 이 데이터의 품질과 신뢰성이 의도한 데이터 응용으로 충분합니까? 설명 해주십시오.
		2A.3 데이터 편향성/가설	2A.3.1 데이터에는 어떤 가설과 편향이 반영되어 있습니까? 알고리즘의 결과물에 미치는 영향이 어떻게 수정 및 극복되거나 완화됩니까? 2A.3.2 학습 데이터가 사용되는 경우: 데이터가 알고리즘이 사용될 맥락을 대표하는가?
		2A.4 보안 및 보관	2A.4.1 데이터가 충분히 안전합니까? 입력 데이터와 출력 데이터를 구분하여 작성하십시오. 2A.4.2 데이터에 대한 접근을 감독하고 있는가? 2A.4.3 보관 규정을 준수하고 있는가?
	2B (알고리즘 - 처리)	2B.1 알고리즘 유형	2B.1.1 2B.1.1 어떤 유형의 알고리즘이 사용됩니까? 다음을 구분하여 작성하십시오. a. 인간이 컴퓨터가 준수해야 하는 규칙을 지정하는 비자기지도 학습 알고리즘 (non-self-learning algorithm) b. 기계 자체가 데이터에서 패턴을 찾는 자기지도 학습 알고리즘(self-learning algorithm) 2B.1.2 이 유형의 알고리즘이 선택된 이유는 무엇입니까? 2B.1.3 이러한 유형의 알고리즘이 질의 1.2에서 서술한 목적을 달성하는 데 가장 적합한 이유는 무엇입니까? 2B.1.4 어떤 대안이 있으며 이 대안들이 덜 적절하거나 유용하지 않은 이유는 무엇입니까?

		2B.2 소유권 및 통제권	2B.2.1 알고리즘이 외부에서 개발된 경우: 알고리즘의 소유권 및 관리 권한에 대해 명확한 합의가 이루어졌습니까? 그 합의 내용은 무엇입니까?
		2B.3 알고리즘 정확성	2B.3.1 알고리즘의 정확도는 얼마입니까? 이 정확도는 어떤 평가 기준에 따라 결정됩니까? 2B.3.2 알고리즘이 사용될 방식에 있어 정확도 수준(질의 2B.3.1)이 적절합니까? 2B.3.3 알고리즘은 어떻게 테스트됩니까? 2B.3.4 편향 복제 또는 증폭의 위험에 대응하기 위해 어떤 조치를 취할 수 있습니까? (예: 다른 샘플링 전략, 기능 수정 등) 2B.3.5 지표의 선택과 가중치의 기초가 되는 가설은 무엇입니까? 그 가설이 유효합니까? 유효하거나 유효하지 않은 이유는 무엇입니까? 2B.3.6 알고리즘에서 얼마나 자주 오류가 있나요? (예: 위양성, 위음성, R제곱(R2) 등에 있어서)
		2B.4 투명성 및 설명가능성	2B.4.1 알고리즘이 무엇을 하는지, 어떻게 하며, 무엇(어떤 데이터)을 기반으로 하는지 명확합니까? 설명하십시오. 2B.4.2 어떤 사람과 집단(내부 및 외부)을 위해 알고리즘을 투명하게 운영할 것이며 어떻게 이를 수행합니까? 2B.4.3 어떤 대상 집단에 대하여 알고리즘이 설명될 수 있어야 합니까? 2B.4.4 알고리즘의 작동이 질의 B.4.3에서 파악된 대상 집단에게 충분히 이해할 수 있는 방식으로 설명될 수 있습니까?
3부: 어떻게 하는가?		3.1 알고리즘 결과에 기반한 의사결정	3.1.1 알고리즘의 결과 또는 결과물은 어떻습니까? 어떤 의사결정이 이를 기반으로 이루어집니까?

	3.2 의사결정에서 인간의 역할	3.2.1 알고리즘의 결과물에 기반한 의사결정에서 인간은 어떤 역할을 합니까? 알고리즘 결과물을 기반으로 하여 책임 있는 의사결정을 내릴 수 있는 역량은 어떻게 부여됩니까? 3.2.2 필요한 경우 알고리즘을 관리, 검토 및 조정할 수 있는 충분한 자격의 직원을 갖추고 있습니까? 앞으로 그럴 예정입니까?
	3.3 알고리즘의 효과	3.3.1 시민들에게 알고리즘 사용의 효과는 무엇이며 알고리즘을 기반으로 의사결정을 내릴 때 '인간의 조치'는 어떻게 고려할 예정입니까? 3.3.2 낙인, 차별, 기타 시민에게 해롭거나 부정적인 영향의 위험은 무엇입니까? 이러한 문제를 어떻게 해결하거나 완화할 수 있습니까? 3.3.3 예상되는 효과가 알고리즘의 개발/배포를 촉발한 문제를 해결하는 데 어떻게 기여하고(질의 1.1 참조), 제안된 목적을 달성하는 데 어떻게 기여할 예정입니까?(질의 1.2 참조) 3.3.4 예상되는 효과는 제공되는 가치와 어떤 관련이 있습니까?(질의 1.3 참조) 특정 가치를 훼손하는 위험은 어떻게 처리됩니까?
	3.4 절차	3.4.1 알고리즘을 기반으로 의사결정을 내릴 때 어떤 절차가 있습니까? 3.4.2 다양한 관련 행위자(행정적 및 정치적 책임이 있는 사람, 시민)가 의사결정에 어떻게 참여합니까? 3.4.3 이러한 절차가 좋은 거버넌스, 좋은 행정 및 필요한 경우 법적 보장의 요구사항을 충족하도록 어떻게 보장하고 있습니까?
	3.5 맥락	3.5.1 시기/기간: 알고리즘은 언제 사용됩니까? 얼마나 오래 사용됩니까? 3.5.2 영역: 알고리즘은 어디에 사용됩니까? 특정 지리적 영역 하에 있습니까? 알고리즘이 특정 집단의 사람 또는 사례와 관련이 있습니까? 3.5.3 맥락적 요인이 변경되거나 알고리즘이 개발된 맥락과 다른 상황에서 사용되는 경우에도 알고리즘을 계속 사용할 수 있습니까?

	3.6 소통	3.6.1 알고리즘 도입의 목적 및 맥락에 비추어 볼 때 알고리즘 운영에 대해 얼마나 개방할 수 있습니까? 3.6.2 알고리즘 사용에 대해 어떻게 소통할 계획입니까? 3.6.3 알고리즘 결과물이 예를 들어 도표, 그래프 또는 대시보드로 시각화됩니까? 만약 그렇다면, 시각화 또는 묘사의 형태가 알고리즘 결과물을 올바르게 대표합니까? 다양한 사용자 집단이 이 시각화를 이해하기에 쉽습니까?
	3.7 평가, 감사 및 보장	3.7.1 알고리즘의 평가, 감사 및 보장을 위해 적절한 도구가 제공되었습니까? 3.7.2 알고리즘에 대해 적절하게 설명할 수 있는 충분한 기회가 있습니까? 3.7.3 감사인과 규제기관이 정부의 알고리즘 사용에 (공식적인) 결과를 첨부할 수 있는 경우로는 어떤 것이 있습니까? (예: 결정에 대한 피드백, 권고 사항, 예산 결정 등)
4부: 기본권 로드맵	4.1. 기본권	4.1.1 사용되는 알고리즘의 영향을 받는 기본권이 있습니까?
	4.2. 구체적인 법률	4.2.1 기본권 침해에 대해 특별한 법률 조항이나 기준이 적용됩니까?
	4.3. 심각도 정의	4.3.1 기본권이 알고리즘에 의해 얼마나 심각하게 영향을 받습니까?
	4.4. 목적	4.4.1 알고리즘을 사용하여 추구하는 목적은 무엇입니까?
	4.5. 적합성	4.5.1 사용할 알고리즘이 설정된 목적을 실현하는 데 적합한 수단입니까?
	4.6 필요성 및 보충성	4.6.1 이 특정 알고리즘을 사용하는 것이 이 목적을 달성하는 데 필요하며 이를 위해 사용할 수 있는 다른 수단 또는 완화 조치가 없습니까?
	4.7 균형성 및 비례성	4.7.1 알고리즘을 사용하면 추구하는 목적과 침해될 기본권 사이에 합리적인 균형이 이루어집니까?