(사) 정보인권연구소 세미나

인공지능과 빅테크
인권영향평가 현황과 과제

오병일
진보네트워크센터 대표
정보인권연구소 연구위원

연구수행기관 : 한동대학교 산학협력단

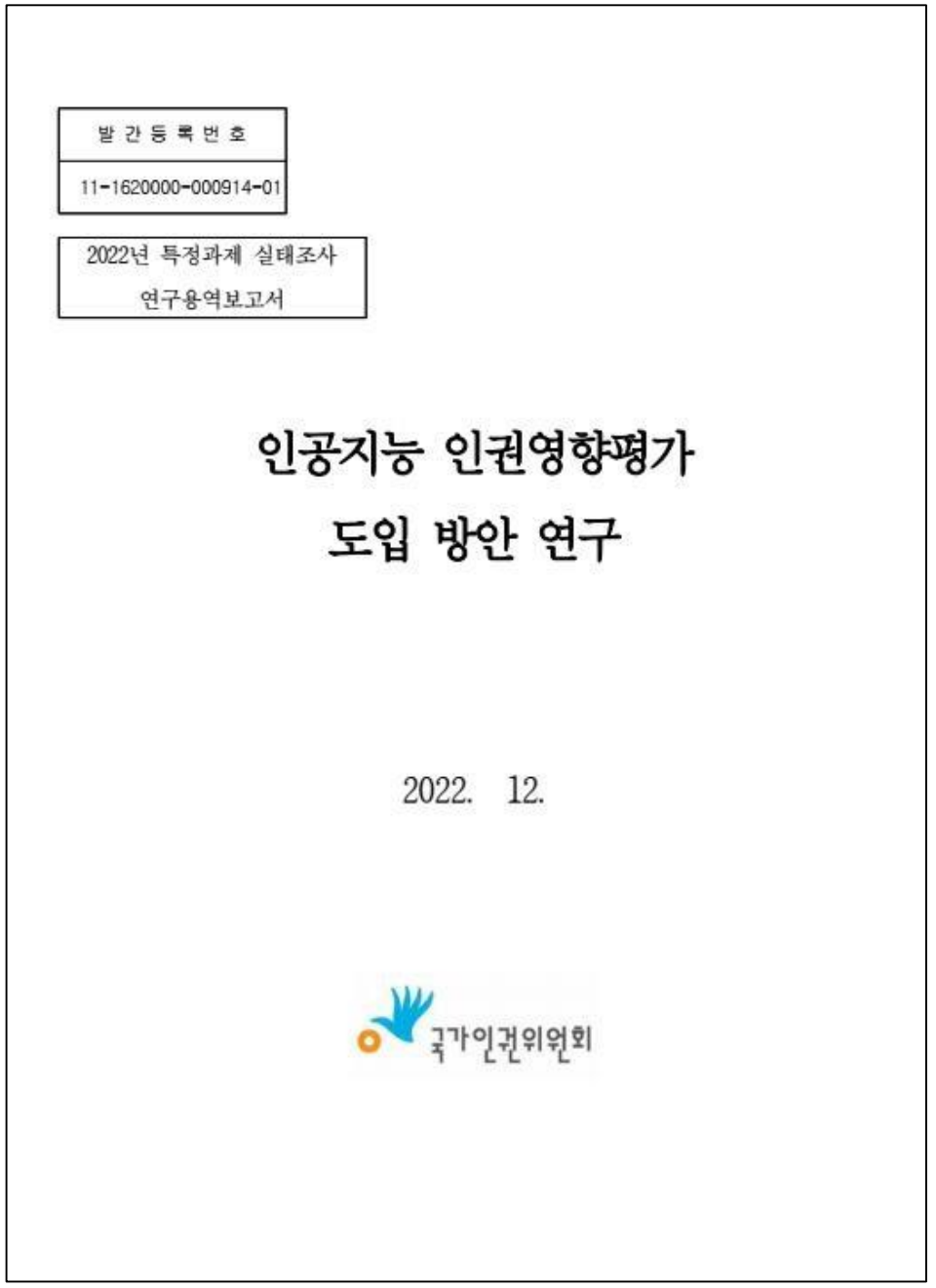
연구책임자 : 유승익 (한동대학교 연구교수)

공동연구원 : 김병욱 (해우법률사무소 변호사)

: 오병일 (진보네트워크센터 대표)

: 오정미 (법무법인 이공 변호사,
사단법인정보인권연구소 연구위원)

보조연구원 : 안영선 (진보네트워크센터 활동가)



인공지능 인권영향평가란?

“(인공지능에 의한) 많은 추론과 예측은, 사람들의 자율성과 자신의 정체성에 대한 세부사항을 확립할 권리를 포함하여, 프라이버시권의 향유에 깊은 영향을 미친다. 이는 또한 사상과 의견의 자유에 대한 권리, 표현의 자유, 공정한 재판 관련 권리 등 다른 권리에도 많은 문제를 야기한다. ”

“인공지능 시스템의 설계, 개발, 배치, 판매, 구입, 운영의 수명주기 전반에 걸쳐 체계적으로 인권실사를 수행한다. 그 인권 실사의 핵심 요소는 정례적이고 포괄적인 인권영향평가여야 한다. “

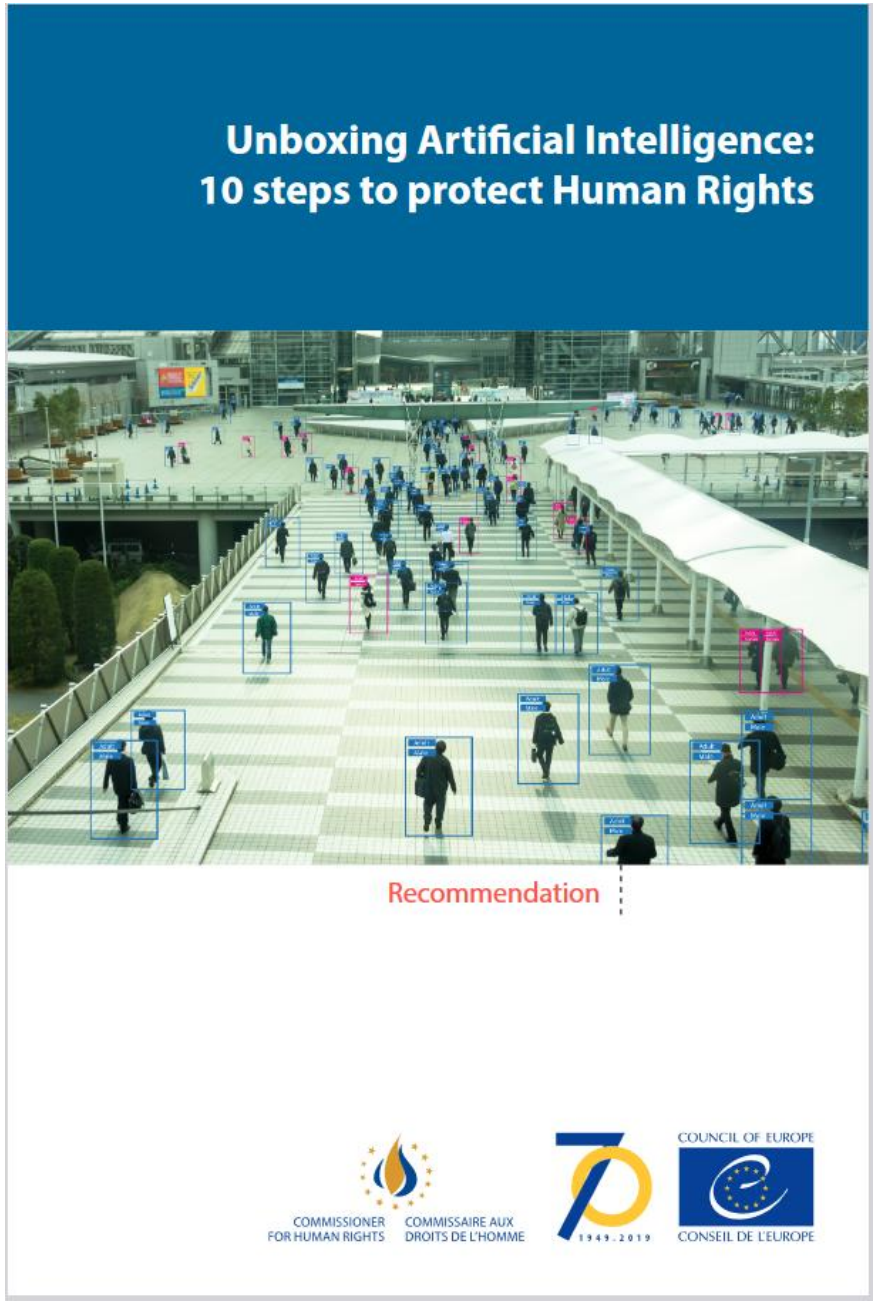
- 유엔 인권최고대표 <디지털 시대 프라이버시권 (2021)> -

"회원국은 공공기관이 구입, 개발 또는 배치하는 인공지능 시스템에 대해 인권영향평가 실시 절차를 도입하는 법적 체계를 수립해야 한다."

"회원국은 공공기관과 민간 기업이 개발, 배치, 사용하는 인공지능 시스템의 인권 준수에 대한 독립적이고 효과적인 감독을 위해 입법 체계를 수립해야 한다. "

유럽평의회 인권위원장

<인공지능 블랙박스 개봉: 인권 보호를 위한 10단계>



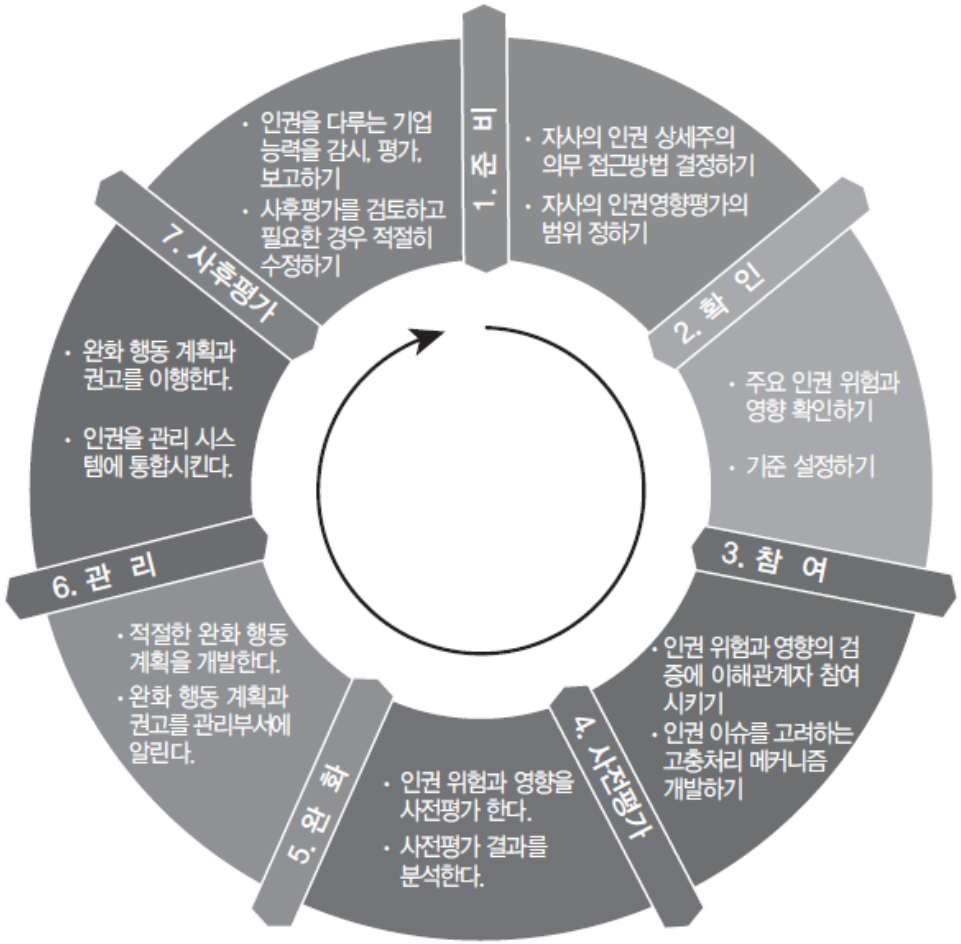
“39. 국가는 인공지능의 개발과 활용에 있어서 인권침해와 차별의 가능성 및 정도, 영향을 받는 당사자의 수, 사용된 데이터의 양 등을 고려하여 공공기관 및 민간기업을 대상으로 인권영향평가를 실시하여야 합니다.”

국가인권위원회

<인공지능 개발과 활용에 관한 인권 가이드라인> (2022.4)

인권영향평가 (Human Rights Impact Assessment)

- 사업 과정, 정책, 입법, 프로젝트 등이 인권에 미치는 영향을 측정하고 평가하는 도구
- 2011년, 기업과 인권에 관한 이행지침 (UN Guiding Principles on Business and Human Rights, UNGPs)
  - 보호, 존중, 구제 프레임워크 : 국가의 인권보호의무, 기업의 인권존중 책임, 효과적인 구제 수단에 대한 접근
  - 부정적 인권 영향을 식별하고 방지하고 완화하며 어떻게 그에 대처하는지를 설명하기 위해서 기업은 인권 실사를 수행해야 함.
  - 인권영향평가는 인권 실사의 핵심 도구



국내 인권영향평가

- 2021.4. 국가인권위원회, '인권 기본 조례 제·개정 권고'와 함께 <인권 기본조례 표준안> 발표
  - 지자체의 인권 관련 조례 제정, 인권영향평가 제도 규정
- 국가인권위, 공공기관 경영평가제도의 일환으로 인권경영 권고
  - 공공기관 인권경영 매뉴얼(2018)에서 인권영향평가 규정.
  - 기관(기업)운영 / 주요사업 인권영향평가로 구분.
- 2025년 환경·사회·지배구조(Environment, Social, Governance, 일명 'ESG') 공시 의무화 영향으로 인권영향평가 도입 움직임

국내 인공지능 평가도구

- 개인정보보호위원회, 인공지능(AI) 개인정보보호 자율점검표 (2021.5)
- 과학기술정보통신부, 신뢰할 수 있는 인공지능 실현전략 (2021.5) 발표
  - 「지능정보화 기본법」 제56조에 따라 인공지능 영향평가를 실시할 방침 표명
- 과학기술정보통신부, 한국정보통신기술협회(TTA), <2022 신뢰할 수 있는 인공지능 개발 안내서 (안)> (2022)
- 금융위원회, <금융분야 인공지능(AI) 가이드라인> (2021.7) 시행, <금융분야 AI 개발·활용 안내서> 발표(2022.8)
- 서울시교육청, 인공지능(AI) 공공성 확보를 위한 현장 가이드라인 (2021.9)

인공지능 영향평가 해외 사례

유럽연합, 신뢰할 수 있는 인공지능 평가 목록 (2020.7)

- 자율점검 : ① 인간행위자와 감독, ② 기술적 견고성과 안정성, ③ 프라이버시 및 데이터 거버넌스, ④ 투명성, ⑤ 다양성, 차별 금지, 공정성, ⑥ 사회·환경적 복지, ⑦ 책무성 등 윤리지침 요구사항 준수 여부 점검을 위한 질의
- 기본권 영향평가 수행 제안
- 인공지능법(안) (2021.4)으로 발전

캐나다 알고리즘영향평가

- 자동화된 의사결정 훈령 제정(2019) : 알고리즘 영향평가 적용
- 공공기관이 의사결정에 사용하는 인공지능 시스템에 적용
- 위험기반 접근법 : 위험 수준을 4단계로 구분, 위험 수준이 높을수록 높은 요구사항 적용
- 해당 기관이 수행, 이해관계자 참여, 평가결과 공개

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.

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]

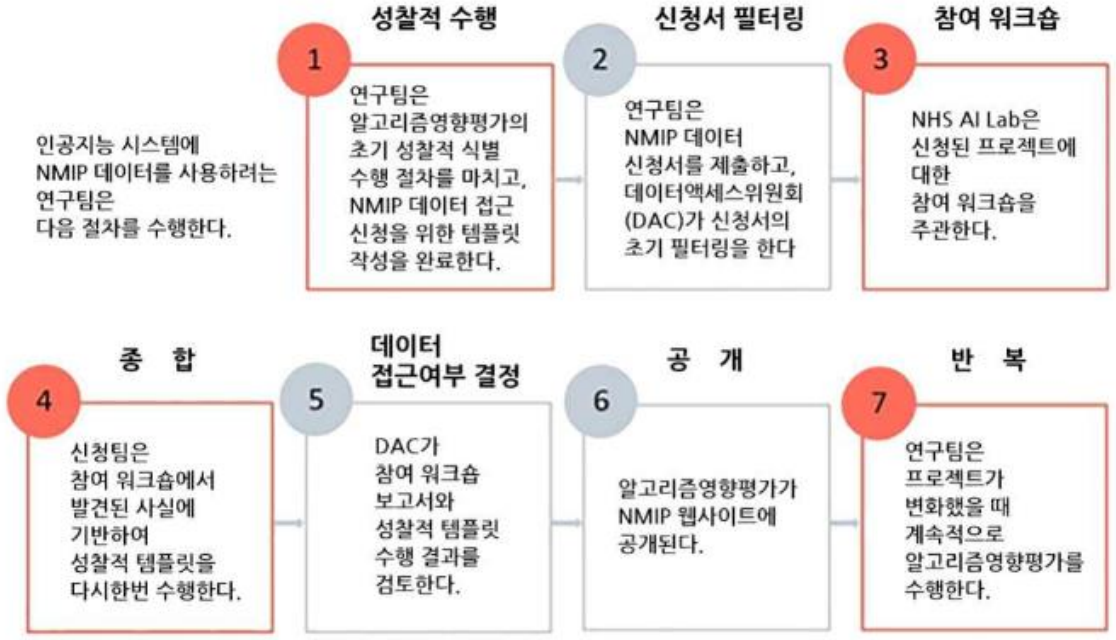
☐ 예
☐ 아니오

영국 인공지능 조달지침 (2020.6)

- 인공지능 공공조달을 위한 10대 원칙 제시
- 조달절차 개시 전, 데이터 평가 실시
- 조달절차 개시단계, 인공지능 배치의 편익과 위험에 대한 영향평가 실시
  - AI 시스템에 대한 사용자 요구사항과 그 공익
  - AI 시스템의 인적 및 사회 경제적 영향 - 이는 AI가 사회적 가치 편익을 제공할 수 있도록 보장함
  - 기존의 기술적, 절차적 환경에 미친 결과
  - 데이터 품질 및 부정확하거나 편향될 가능성
  - 의도하지 않은 결과가 나올 가능성
  - 지속적인 지원 및 유지보수 요구사항을 비롯해 전체 생애주기에 대한 비용적 고려사항

영국, NMIP 알고리즘 영향평가

- 보건의료 분야 국가의료이미지플랫폼(National Medical Imaging Platform, NMIP)에 특화된 알고리즘 영향평가 도입
- 에이다 러브레이스 연구소가 영국 NHS AI Lab의 지원을 받아 2021년 개발



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

미국 알고리즘 책무성법(안)

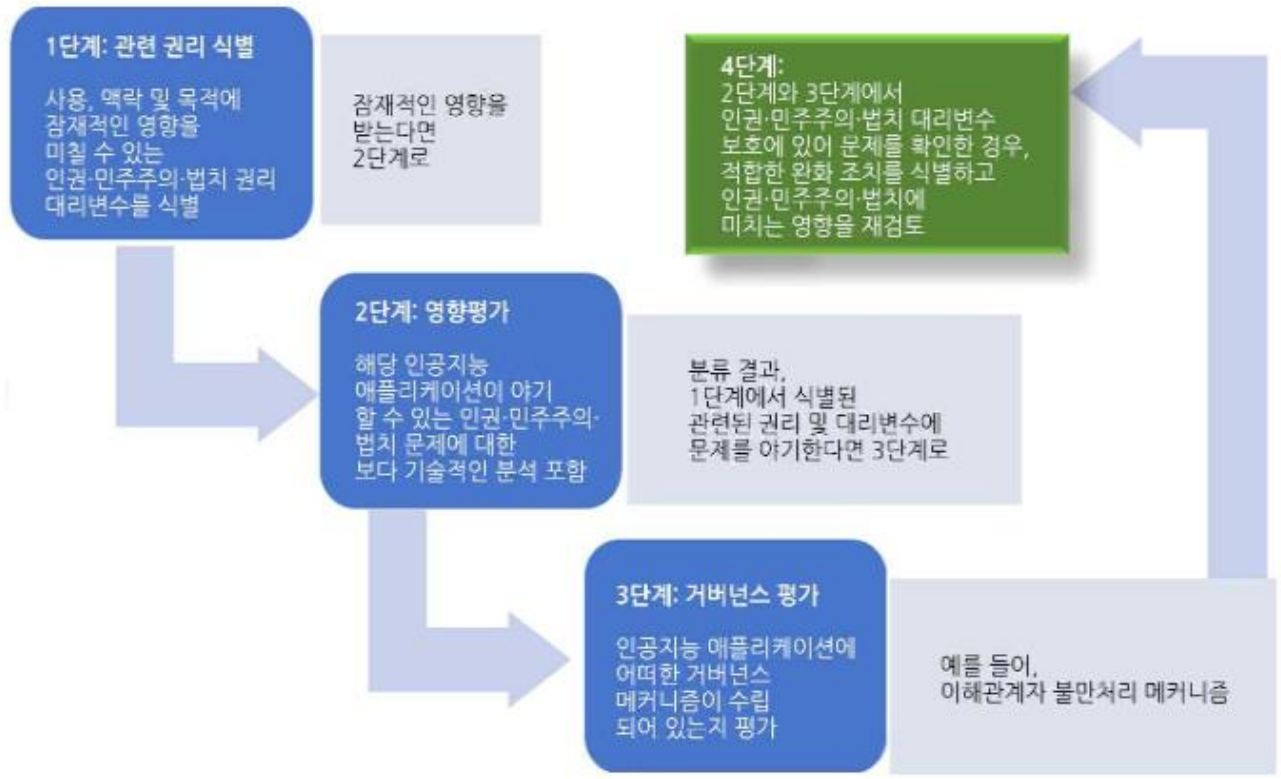
- 2022.2. 상하원 동시 발의
- 연방거래위원회(FTC)가 소관하는 일정 규모 이상의 기업들을 대상으로 자동화된 의사결정 시스템 또는 증강된 중요 의사결정 프로세스, 이들이 소비자에게 미치는 영향에 대하여 지속적으로 연구·점검하는 '영향평가'(§2(12)) 실시 의무화

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

유럽평의회, 인권·민주주의·법치 영향평가

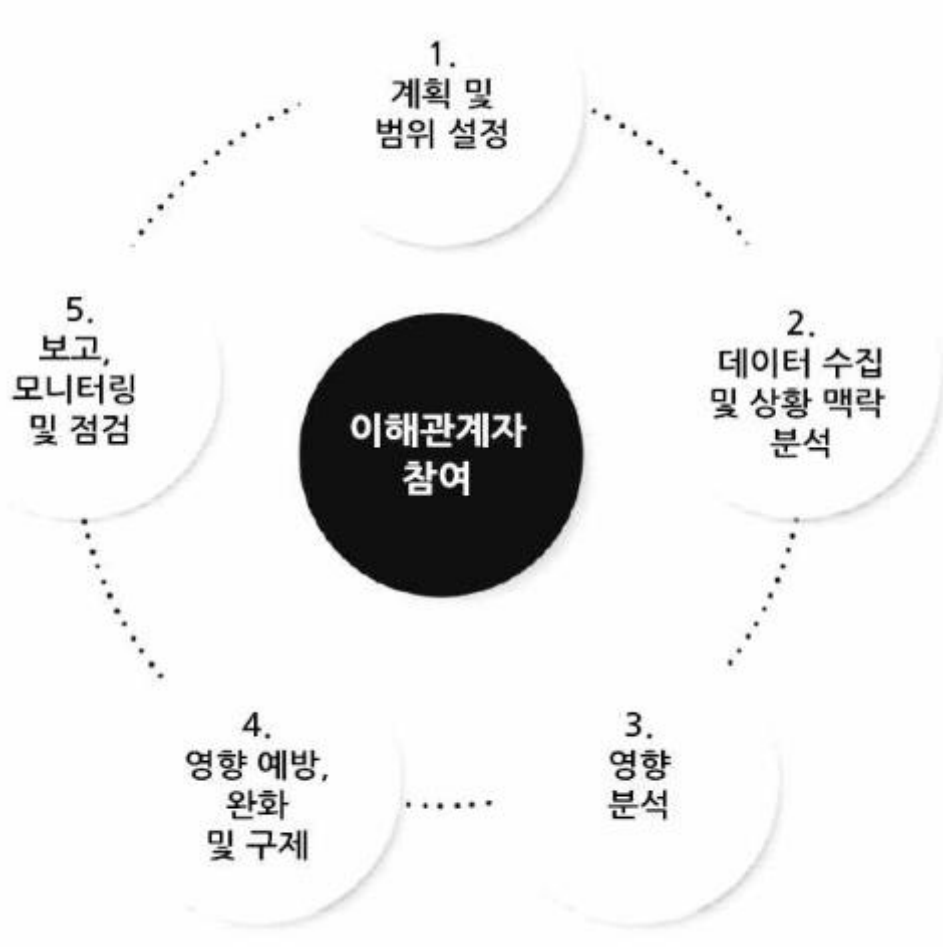
- Human Rights, Democracy, and Rule of Law Impact Assessment of AI, HRDRIA (2021.5.)
  - 인공지능 시스템이 운영되는 지정학적, 사회적, 경제적 맥락 고려
  - 인공지능 시스템 기반 기술의 특성 고려 :
    범위, 신뢰성, 추적 가능성, 설명가능성 등
  - 이해관계자의 참여



덴마크 디지털활동 인권영향평가

- AI 등 디지털활동, 제품 및 서비스가 야기, 기여, 관련된 위험의 특성을 평가하고 해결하기 위한 지침
- 인권영향평가 절차 및 내용의 10개 핵심 요소
  - 절차 : 참여, 차별금지, 역량 강화, 투명성, 책무성
  - 내용 : 인권 기준, 영향 범위, 심각도 평가, 완화조치, 구제수단 접근
- 심각도 평가 지표 제시 : 범위, 규모, 회복불가능성

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.
	영향 영역 전체 인구의 10% 이상 또는 식별된 집단의 10-50%	B	
	영향 영역 전체 인구의 5% 이상 또는 식별된 집단의 10% 미만	C	

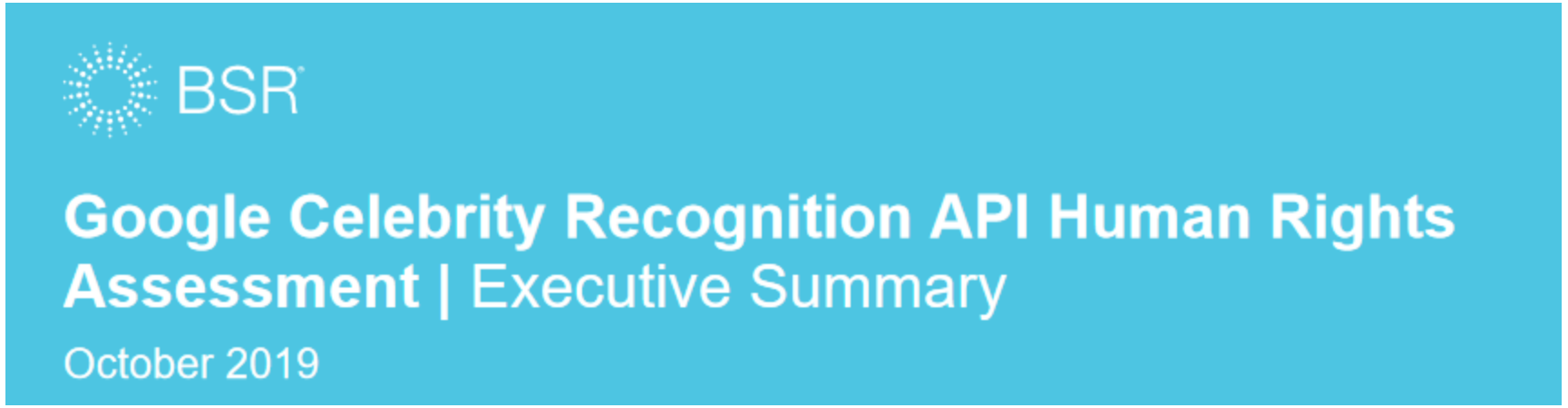


네덜란드 기본권 알고리즘영향평가 (2022)

- 인공지능 사회복지급여 부정수급탐지시스템(SyRI)에 대한 헤이그 지방법원의 위법 결정 (2020)
- 인공지능 시스템을 개발하거나 도입하는 초기 단계에서 논의하고 해결하여야 할 인권 쟁점에 대한 질의로 구성

구글, 유명인 인식 API 인권영향평가

- 영상 콘텐츠에서 유명인을 식별하는 API에 대한 인권영향평가 : 기업 협회인 BSR이 수행 (2019)
- 유엔 기업과 인권 이행지침 기반 : 이해관계자 협의, 독립적 전문가 자문, 취약 그룹 고려
- API를 전문 영상 콘텐츠로 제한, 당사자 요청에 의한 옵트아웃 정책, API 사용 고객 리스트 구축 등 권고



페이스북, 국가별 인권영향평가

- BSR, 미얀마 페이스북 인권영향평가 수행 (2018)
  - 페이스북이 미얀마 지역 분열과 폭력의 조장을 방지하는데 충분한 역할을 하지 못했다고 결론, 시정 조치 권고
  - 하버드 케네디스쿨의 인권정책을 위한 카센터(CARR Center) 연구자, 페이스북 뉴스피드 알고리즘의 인권 영향에 대한 평가가 미흡하다고 비판
- 스리랑카, 인도네시아, 캄보디아에서의 인권영향평가 발표 (2020)
- 인도에서의 인권영향평가 : 인권영향평가의 독립성 침해, 영향평가 보고서 비공개 논란

인공지능 인권영향평가(안)

인공지능 인권영향평가 도구(안) 개요

- 대상 : 고위험 인공지능 + 공공기관이 도입하는 인공지능
- 시기 : 사전(인공지능기술 개발 또는 도입에 관한 구상이 구체화된 시점) + 정기적인/필요할 경우 사후 재평가
- 수행 주체 : 내부 혹은 외부의 독립적인 조직
    - 예를 들어, 인공지능 윤리, 인권 경영, ESG 경영 등을 담당하는 부서, 혹은 독립성과 인권 및 인공지능 전문성을 갖춘 제3의 기관

인공지능 인권영향평가 절차

인공지능 인권영향평가 도구(안)의 구성 방식

- 각 단계별로 점검해야 할 질의와 질의 취지에 대한 설명으로 이루어짐
- 질의에 대한 답변은 주관식인 경우도 있고, 객관식인 경우도 있음.
- 객관식인 경우에도 관련 내용에 대한 설명을 제공하도록 함. (이후 영향평가결과 공개)
- 각 챕터마다 질의와 관련된 국내외 사례에 대한 정보 제공

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.

【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.

인공지능 인권영향평가 도구(안)의 주요 내용

- 【1단계 : 계획 및 준비】
- 가. 인권영향평가 계획
  - 평가 대상 인공지능 시스템에 대한 설명 및 이해관계자 파악을 위한 질의
- 나. 조사
  - 인공지능 시스템, 도입될 시공간적 맥락,
    이해관계자와의 협의 관련 자료 확보를 위한 질의

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

인공지능 인권영향평가 도구(안)의 주요 내용

- 【2단계 : 분석 및 평가】
- 가. 인공지능 기술과 관련된 영향 분석 및 평가
  - (1) 개인정보보호 : 개인정보보호법 준수 여부 확인
  - (2) 데이터 : 데이터셋의 정확성, 다양성 등 확인
  - (3) 알고리즘의 성능과 신뢰성 : 성능 측정 지표 및 적절성 확인
  - (4) 차별금지
  - (5) 설명가능성과 투명성
  - (6) 자동화 정도와 인간의 개입 : 인간의 개입 가능성 확인
  - (7) 보안
  - (8) 접근성 : 인터페이스의 보편적 설계 확인
  - (9) 라이선스 : 이용자의 수정 가능성 확인

인공지능 인권영향평가 도구(안)의 주요 내용

- 나. 인권에 미치는 영향 및 심각도
  - (1) 영향을 받는 인권
  - 헌법 및 국제인권규범에 근거
  - 침해 가능성이 있는 인권의 예시 제공
  - (2) 인권에 미치는 영향의 심각도
  - 영향의 범위, 규모, 회복 불가능성 파악

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

인공지능 인권영향평가 도구(안)의 주요 내용

- 【3단계 : 개선 및 구제】
- 가. 방지
  - 인권 침해 방지 조치 검토
- 나. 완화
  - 인권 침해 완화 조치 검토
- 다. 구제
  - 권리구제 조치 검토
- 라. 이해관계자와의 의견수렴 및 협의
  - 개선 및 구제 조치에 대한 이해관계자 협의 여부

인공지능 인권영향평가 도구(안)의 주요 내용

- 【4단계 : 공개 및 점검】
- 가. 인공지능 시스템의 주요 요소의 공개
- 나. 인권영향평가 결과 공개
- 다. 인공지능 시스템에 대한 모니터링
- 라. 인권영향평가에 대한 점검
  - 평가 절차 및 효과에 대한 검토
- 마. 인권영향평가의 재수행

인공지능 인권영향평가 도구(안)에 대한 주요 문제제기

- 평가 대상인 고위험 인공지능이 무엇인가
- 인권영향평가가 또 하나의 규제가 되는 것이 아닌가 : 질의는 검토 사항인가 권고 사항인가
- 법제화없는 인권영향평가가 실효성이 있을까
- 인권영향평가 과정에서 기업 영업비밀 침해 우려
- 평가 수행 주체의 모호함 : 자율점검인가, 외부적인 평가인가
- 다른 평가나 규범과의 관계 : ex) 개인정보 영향평가, 과기부의 <인공지능 개발 안내서> 등

감사합니다