발간등록번호
서교연 2020-27

2020 위탁연구 보고서

공교육에 적용되는 인공지능 알고리즘의 공공성 확보방안 연구

서울특별시교육청
교육혁신과

2020 위탁연구 보고서

공교육에 적용되는 인공지능 알고리즘의 공공성 확보방안 연구

2021. 3.

주관연구기관 : 사단법인 정보인권연구소

연구책임자 : 김기중 (법무법인 동서양재 변호사)

공동연구원 : 임완철 (경상대학교 교수)

장여경 (사단법인 정보인권연구소 상임이사)

서울특별시교육청

교육혁신과

이 연구는 서울특별시교육청 교육혁신과 연구지원비로 수행되었으나, 본 연구에서 제시된 정책대안이나 의견 등은 서울특별시교육청 교육혁신과의 공식 의견이 아니라 본 연구팀의 견해임을 밝혀 둡니다.

연구결과 요약

1. 연구 목적 및 필요성

학생의 개인정보 보호와 데이터 처리 과정의 투명성 확보를 위하여 공교육 적용 인공지능 알고리즘에 대한 관리방안이 마련될 필요성이 있다. 이에 본 연구는 공교육에 적용되는 인공지능 알고리즘에 설명가능성, 투명성, 책무성, 책임성 등의 원칙을 적용하기 위한 정책 방안 마련을 목적으로 하였다.

2. 연구 방법

본 연구는 문헌 연구를 통하여 인공지능 알고리즘 관련 국내외 규범을 검토하였다.

3. 연구 내용

본 연구의 Ⅱ장에서는 인공지능 윤리, 공공기관 인공지능 윤리, 관련 법령 등 인공지능 알고리즘 관련 일반 규범을 검토하였다.

Ⅲ장에서는 캐나다 정부 〈자동화된 의사결정에 대한 지침〉, 뉴질랜드 정부 〈위험성 매트릭스〉, 독일 정부 데이터윤리위원회 〈알고리즘 시스템 위험도 피라미드〉, 유럽연합 〈위험기반 접근법〉과 싱가포르 정부 〈위험평가 매트릭스〉 등 해외 여러 국가에서 인공지능 알고리즘 시스템의 위험성을 평가하고 위험 등급별 관리를 도입하였거나 추진 중인 현황을 검토하였다.

이어 Ⅳ장에서는 그밖의 인공지능의 모범 정책으로 인공지능 영향평가, 투명한 정보공개와 참여 보장, 정보주체의 권리 보장 정책 등에 대하여 검토하고, 국내 적용 방안을 제시하였다.

마지막으로 Ⅴ장에서는 이상의 연구 검토 결과를 토대로 공교육에 적용되는 인공지능 알고리즘의 공공성을 확보하기 위한 종합적 제언을 도출하였다.

4. 연구 결과

학습자의 성장을 최우선 원칙으로 하는 공교육 적용 인공지능 알고리즘 원칙을 수립할 필요성이 확인되었다. 공교육 학교시스템에 포함되는 학생은 대부분 신체적으로나 정신적으로 성인에 이르기까지 발달하는 과정에 있는 인간이라는 특수성을 고려하여 공교육에 적용되는 인공지능 알고리즘 원칙을 고안해야 한다. 또한 인공지능 알고리즘에 대한 개인적/사회적/기술적 차원의 통제력 확보 체제를 구축할 필요가 있다.

이에 인공지능 알고리즘의 공교육 적용을 위한 거버넌스 방안을 제안하고자 한다. 첫째, 인공지능이 교육에 미칠 영향 평가 (인공지능 영향평가) 체제의 구축이 필요하다. 둘째, 인공지능 알고리즘의 투명성 확보(정보공개와 참여) 체제를 구축할 필요가 있다. 셋째, 정보주체로의 성장을 지원할 역량 강화 프로그램의 구축이 필요하다. 넷째, 공교육에 적용되는 인공지능 알고리즘의 공공성 확보를 위한 거버넌스 체제를 구축할 필요가 있다.

5. 정책제언

① 공교육 인공지능 알고리즘 원칙(헌장), 영향평가 도구와 프로세스, 조달 가이드라인 등 개발

- 모든 학생과 교원에게 이로울 수 있는, 공교육 인공지능 알고리즘 원칙(헌장) 개발
- 공교육 인공지능 알고리즘 영향평가(위험성 매트릭스 등) 도구와 프로세스 개발
- 공교육 인공지능 알고리즘 조달 가이드라인 개발

② 공교육 인공지능 알고리즘 원칙의 실행을 위한 추진기구 설립

- 공교육 인공지능 알고리즘 원칙(헌장) 개발 및 발표
- 공교육 인공지능 알고리즘 영향 평가 도구 개발
- 공교육 인공지능 알고리즘 영향 평가 프로세스 구축
- 공교육 인공지능 알고리즘 영향 평가에 따른 위험성 등급 평가
- 위험성 등급에 따른 이후 처리 절차(테스트베드 상황에서의 시범 적용 등) 정의
- 영향 평가 사후 처리 결과를 공개하고 후속 과정 진행

차 례

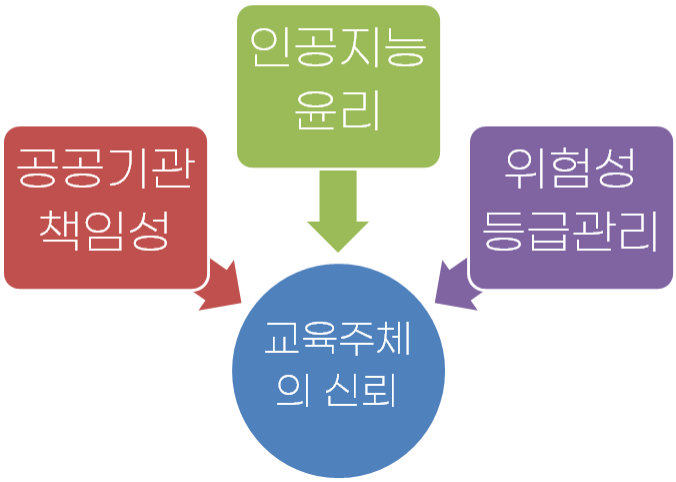
▸알고리즘은 정부가 뉴질랜드와 뉴질랜드인을 더 잘 이해하도록 사용될 수 있다. 이러한 지식들은 정부가 좋은 결정을 내리고 보다 효율적이고 효과적인 서비스를 제공하는 데 도움이 된다. 알고리즘의 사용은 정부 서비스의 집행에 인간의 편견이 작용할 위험을 방지하고 모두를 위한 혜택을 현실화할 수 있다.

▸하지만, 이러한 기회는 새로운 문제를 가져오기도 한다. 예를 들어 세심하게 설계되거나 구동되지 않는 알고리즘은 인간의 편견을 고착화하거나 심하게는 강화할 수도 있다. 투명성과 책무성은 정부가 이러한 도구를 적합한 방식으로 사용한다고 시민들이 신뢰하고 지지할 수 있게끔 하는 데 매우 중요하다.

▸본 헌장[및 알고리즘 위험성 매트릭스]은 정부 기관이 개인정보 보호와 투명성 사이의 올바른 균형을 유지하고, 의도치 않은 편견을 방지하며 와이탕이 조약(Treaty of Waitangi)의 원칙을 반영하여 알고리즘을 신중하게 사용하겠다는 약속이다.

- 뉴질랜드 정부 알고리즘 헌장 (2020. 7)

공교육과 인공지능



I. 서문

ㅇ 공공부문에 도입되는 인공지능 알고리즘은 때로 국민들에게 법적, 또는 이와 유사하게 중대한 영향을 미치는 자동화된 의사결정[1]에 이를 수 있다는 점에서 엄격한 책임성 보장이 요구됨

ㅇ 초중고등학교 등의 공교육에 도입되는 인공지능 알고리즘은 미래 시민이 될 아동학생의 발달에 미치는 영향이 적지 않을 것으로 예상됨. 특히 공적인 평가 및 의사결정을 지원하거나 이에 직접 작용하는 인공지능 알고리즘은 각 학생의 미래에 중대한 영향을 미칠 가능성을 배제할 수 없음

【사례】 교육평가 분야 인공지능 알고리즘의 차별 논란[2]

▸영국 시험감독청(Ofqual)은 2020년 코로나19로 대학수학능력시험에 해당하는 A레벨 시험을 취소하는 대신 인공지능 알고리즘을 통해 학생 성적을 부여함. 이 알고리즘은 각 학생의 A레벨 예비시험과 학교 과제 점수, 교사의 예상치 등을 바탕으로 성적을 산출하고 소속 학교의 역대 학업능력을 고려하여 가중치를 부과함

▸그러나 평가 결과 부유한 지역 학생이 높은 점수를 받은 반면 가난한 지역 학생은 상대적으로 차별을 받은 것으로 나타남. 인공지능이 불평등을 강화한다며 영국 전역에서 시위가 벌어지고 이 사태로 교육부 담당 공무원과 시험감독청장이 사임함.

2020년 8월 영국 교육부 장관과 시험감독청장은 A레벨 알고리즘 성적을 철회한다고 밝히고 교사가 제출한 예상치에 따라 새 성적을 부여한 후 “대학에는 당국과 교사가 산출한 성적 중 더 높은 수치를 제공하겠다”고 밝힘

1) “법적, 또는 이와 유사하게 중대한 영향을 미치는 자동화된 의사결정”의 경우에는 유럽연합 개인정보보호법(GDPR) 등에서 최근 규제 대상으로 포섭되고 있는 법률적 개념임

2) 가디언 관련 보도 <https://www.theguardian.com/education/2020/aug/13/who-won-and-who-lost-when-a-levels-meet-the-algorithm>; 한겨레21 관련 보도

ㅇ 이에 공공기관, 특히 공교육 기관은 인공지능 윤리를 준수함으로써 신기술이 국민에게 가져올 혜택을 헌법이 보호하는 국민의 인권과 조화시키는 정책을 구상하고 실행할 필요가 있음

- 유럽연합 집행위원회는 정부 및 공공기관과 시민 간의 관계를 인공지능이 변화시키고 형성하고 있다고 보고, 유럽이 지향하는 신뢰가능 인공지능의 혁신을 공공기관이 이끌어야 한다고 지적함[3)]

- 스탠퍼드대학교-뉴욕대학교 공동연구는 미국 정부가 인공지능 시스템을 어떻게 도입하였는지 분석한 후 "성능이나 알고리즘 편향성으로 인해 정부와 시민들 사이의 신뢰가 떨어질 수 있다."고 경고함[4)]

ㅇ 특히 공교육 분야 인공지능 알고리즘의 투명성, 공정성, 책임성 부족은 사회적 논란과 공교육에 대한 신뢰 저하; 교사, 학생, 학부모 등 교육주체들의 교육정책을 향한 불신을 초래하여 교육정책 의사결정 과정에서의 불확실성을 증가시킬 우려가 있음

【사례】 교육청 교사평가 인공지능 알고리즘과 투명성 논란[5)]

▸미국 휴스턴 교육청은 공립학교 교사의 고용을 결정함에 있어 민간회사 인공지능 비밀 알고리즘의 평가를 따름. 휴스턴 교사연맹은 적법절차 위반을 주장하는 소송을 제기함

▸2017년 5월 휴스턴 지방법원은 "공공기관이 매우 중요한 노동 관련 의사결정을 할 때 민간회사의 비밀 알고리즘에 기반한다면, 이는 최소한의 적법절차를 준수하기 어렵다. 따라서 적법절차와 영업비밀을 모두 지키기 위한 적절한 해결책은 비밀 알고리즘의 공공 도입을 중단하는 것"이라고 판시함

<http://h21.hani.co.kr/arti/culture/culture_general/49206.html>

3) 유럽연합은 2020년 5월 발표한 <인공지능 기반 서비스 및 솔루션 공공조달의 데이터 윤리 백서>에서 이와 같이 지적함. 이하에서는 '인공지능 공공조달 백서'로 지칭함

4) David Freeman Engstrom, Daniel E. Ho, Catherine M. Sharkey, Mariano-Florentino Cuéllar (2020). "Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies". <https://law.stanford.edu/education/only-at-sls/law-policy-lab/practicums-2018-2019/administering-by-algorithm-artificial-intelligence-in-the-regulatory-state/acus-report-for-administering-by-algorithm-artificial-intelligence-in-the-regulatory-state/#slsnav-report>; 관련 언론보도 <https://venturebeat.com/2020/02/19/only-15-of-ai-federal-agencies-use-is-highly-sophisticated-according-to-stanford-and-nyu-report/>

ㅇ 세계 각국 정부는 공공부문 인공지능 알고리즘이 헌법과 현행 법률 규범에 부합하고 국민 앞에 투명하고 민주적인 책임을 다할 수 있도록 여러 정책적 노력을 경주하고 있음

- 캐나다, 뉴질랜드, 독일 등 각국 정부가 마련했거나 논의 중인 인공지능 알고리즘 관련 규범들은 공공부문 인공지능 알고리즘에 대하여 투명성, 공정성, 합법성 등의 인공지능 윤리 이행을 법규적 수준으로 요구하고 있으며, 그 위험성에 따라 차등적으로 규율하는 정책을 실시 중이거나 준비 중임

- 특히 널리 알려진 유럽연합의 '신뢰가능 인공지능 가이드라인' 및 규제 프레임워크에서는 인공지능의 위험 강도별로 이를 완화하기 위한 적절한 조치가 필요함을 명시함

ㅇ 우리나라에서도 공공과 민간 여러 영역에서 인공지능 제품과 서비스와 관련한 논란이 불거지기 시작함

- 포털 사이트 뉴스 편집 알고리즘에 대한 공정성과 신뢰성에 대한 의문이 제기됨

- 2013년 한맥투자증권이 차익거래 자동매매시스템의 알고리즘 오류로 2분 만에 450억 원의 손실을 입었고 결국 1년 후 파산하는 사고가 발생함

- 2021년 인공지능 챗봇 '이루다'의 혐오 발언, 개인정보 보호법 위반 논란이 크게 불거져 개인정보 보호위원회, 국가인권위원회 등에서 조사를 실시 중임

- 이에 입법조사처는 2021년 2월 보고서에서 정책적 과제로 인공지능 알고리즘 위험성에 따른 차등적 관리 체계 검토를 권고함[6)]

ㅇ 본 연구에서는, 해외 관련 기준과 정책에 대한 검토를 통하여 공교육에 도입되는 인공지능 알고리즘의 위험성을 평가하고 수준별 요건 및 관리 기준을 제시하고자 함

---

5) 판결문 참조 HOUSTON FED. OF TEACHERS v. HOUSTON INDEPENDENT. <https://www.leagle.com/decision/infdco20170530802#>

6) 심용우, 정준화 (2021). "'이루다'를 통해 살펴본 인공지능 활용의 쟁점과 과제". 국회입법조사처 이슈와논점 제1799호(2021. 2. 15).

- Ⅱ장에서는 인공지능 윤리, 공공기관 인공지능 윤리, 관련 법령 등 인공지능 알고리즘 관련 일반 규범을 검토함

- Ⅲ장에서는 캐나다 정부 <자동화된 의사결정에 대한 지침>, 뉴질랜드 정부 <위험성 매트릭스>, 독일 정부 데이터윤리위원회 <알고리즘 시스템 위험도 피라미드>, 유럽연합 <위험기반 접근법>과 싱가포르 정부 <위험평가 매트릭스> 등 해외 여러 국가에서 인공지능 알고리즘 시스템의 위험성을 평가하고 위험 등급별 관리를 도입하였거나 추진 중인 현황을 검토함

- Ⅳ장에서는 그밖의 인공지능의 모범 정책으로 인공지능 영향평가, 투명한 정보공개와 참여 보장, 정보주체의 권리 보장 정책 등에 대하여 검토함

- 마지막으로 Ⅴ장에서는 이상의 연구 검토 결과를 토대로 공교육에 적용되는 인공지능 알고리즘의 공공성을 확보하기 위한 종합적 제언을 도출함

Ⅱ. 인공지능 알고리즘 관련 일반 규범 검토

1. 인공지능 윤리

□ 국제 규범

○ 유럽연합은 2019년 4월 ‘신뢰가능 인공지능 가이드라인’을 채택함. 이 가이드라인은 “모든 시민이 인공지능의 혜택을 누릴 수 있는 인간 중심의 윤리적 목적을 달성”하는 동시에 “신뢰할 수 있는 인공지능 기술의 발전 기준”을 구체적으로 제시함

- 특히 신뢰가능 인공지능의 기반으로 ‘아동’ 등 사회적 약자의 권리 및 정보의 불균형에 대응할 것을 밝힘

【유럽연합】 신뢰가능 인공지능 윤리 가이드라인[7)]

3대 요소	핵심 지침
Ⅰ. 신뢰가능 인공지능의 기반	① 인간 존중을 윤리 원칙으로 준수하는 인공지능 시스템 개발・배포・사용 - 자율성, 위해예방, 공정성 등을 고려 ② 아동・장애인・고용주와 근로자 또는 기업과 소비자 간 권력 및 정보의 불균형에 대응 - 인공지능 기술이 불이익을 주거나 기술 혜택으로부터 배제 가능성이 있는 취약한 집단 배려 ③ 인공지능 기술이 개인과 사회에 상당한 혜택과 이익을 주지만 특정 위험도 초래할 가능성에 주의 - 위험 강도에 따라 이를 완화하기 위한 적절한 조치 필요
Ⅱ. 신뢰가능 인공지능의 실현	① 인적 관리 및 감독 - 인공지능 시스템은 인간의 기본권을 보장하고 자율성을 저해하지 않는 평등한 사회를 구현해야 함 ② 기술적 견고성 및 안전성

7) European Commission (2019). “Ethics guidelines for trustworthy AI”.

	- 인공지능 시스템 알고리즘은 모든 생애주기에서 오류와 오작동 등 처리가 가능한 안전성을 갖추어야 함 ③ 사생활 보호 및 데이터 거버넌스 - 시민은 자신의 데이터(개인정보)를 완전히 삭제할 수 있어야 하며 관련 데이터가 인간에게 해를 입히거나 차별해서는 안 됨 ④ 투명성 - 인공지능 시스템은 설명가능해야 함 ⑤ 다양성, 차별 금지 및 공정성 - 인공지능 시스템은 모든 범위의 인간 능력과 기술 및 요구 사항을 고려하고 접근성을 보장해야 함 ⑥ 사회 복지 및 환경 복지 - 인공지능 시스템은 긍정적인 사회 변화를 주도하고 지속가능한 성장을 이끄는데 활용되어야 함 ⑦ 책임성 - 인공지능 시스템과 그 결과에 대한 책임, 그 책임을 보장하기 위한 구조적 장치를 마련해야 함
Ⅲ. 신뢰가능 인공지능의 평가	Ⅱ단계에서 요구 사항을 실제 사례에 적합하게 적용할 수 있는 기틀 마련 - 인공지능 시스템에 대한 요구 사항과 솔루션 평가 기준 확립 - 인공지능 시스템의 생애주기 전반에 걸쳐 성과를 개선하고 이에 대한 이해관계자 참여 등

*요약번역: 한국과학기술기획평가원(KISTEP), 일부수정.

- 유럽연합 집행위원회는 가이드라인 채택 후 <인공지능 백서(2020. 2)>[8], <인공지능 공공조달 백서(2020. 5)>[9]를 발표하며 인공지능 거버넌스 프레임워크를 제시함

8) European Commission (2020a). “WHITE PAPER: On Artificial Intelligence - A European approach to excellence and trust”.
9) European Commission (2020b). “White Paper on Data Ethics in Public Procurement of AI-based Services and Solutions”.

○ 경제협력개발기구(OECD)는 2019년 5월 ‘OECD 인공지능 권고안’을 공식 채택하고 신뢰가능한 인공지능 구현을 위한 5가지 원칙을 제안함[10)]

【OECD】 인공지능 권고안

1 포용 성장, 지속가능 발전, 복지 증진
2 인간중심 가치 지향, 공정성 지향
3 투명성 확보, 설명가능성 확보
4 보안 및 안전성 확보
5 책임성 확보

○ 유엔 인권최고대표실은 2018년 <인공지능 기술과 표현의 자유> 발표문에서 국제인권법에 기반한 인공지능 규제 체제로서 △인권 원칙 △투명성 △인권영향평가 △감사 △개인의 자율성 △고지및 동의 △권리구제 보장을 제안함[11)]

【유엔 인권최고대표실】 인공지능 인권법 규제 체제

▸인권 원칙: 인공지능은 모든 다른 기술과 마찬가지로 국제인권법에 따른 국가의 의무와 민간기업의 책임을 준수하여 설계되어야 하고 개발되어야 하고 도입되어야 함. 기업은 그 표준, 규정, 시스템 설계를 보편 인권 원칙에 맞추어야 함
▸투명성: 인공지능 시스템은 개인에게 적극적으로 공개되어야 하며, 이들이 인공지능 절차에 자신의 데이터를 적용하거나 투여한다는 사실을 이해할 수 있는 방식으로 공개되어야 함. 기업과 정부는 인공지능 가치 체계의 각 측면에 걸쳐 투명성을 수용해야 함. 기업은 개인 이용자에게 인공지능 시스템의 존재 여부, 그 목적, 구성 및 영향에 대해 교육해야 함. 기업은 얼마나 많은 내용이 삭제되고, 얼마나 자주 삭제를 요청받는지, 얼마나 자주 삭제에 대한 이의가 제기되는지를 공개해야 함

10) OECD (2019). “OECD Principles on AI”. <https://www.oecd.org/going-digital/ai/principles/>

11) The Office of the High Commissioner for Human Rights (2018). “Artificial Intelligence Technologies and Freedom of Expression”. <https://www.ohchr.org/Documents/Issues/Expression/Factsheet_3.pdf>.

▸인권영향평가: 정부와 기업은 인공지능 시스템을 면밀히 조사하고 개념에서 구현에 이르기까지 이의를 제기할 수 있는 조치를 취해야 함. 인권영향평가는 인공지능 시스템의 인권 영향 문제를 해결하기 위한 하나의 도구임

▸감사: 인공지능 시스템의 외부적 검토를 촉진하는 것은 엄격하고 독립적으로 투명성을 보장하는 데 중요함

▸개인의 자율성: 인공지능이 개인의 의견 형성 및 보유 역량과 정보 환경에서 접근하고 표현하는 역량을 비가시적으로 대체하거나 조작하거나 방해해서는 안 됨. 개인의 자율성을 존중하는 것은 최소한 이용자가 지식, 선택 및 통제권을 갖도록 보장하는 것을 의미함

▸고지 및 동의: 기업은 플랫폼, 사이트 또는 서비스의 이용에 자사 인공지능이 어떻게 관여하고 있는지를 이용자에게 충분히 알려야 함

▸권리구제: 인공지능 시스템이 인권에 악영향을 미친다면 관련 기업이 이를 구제하는 것이 가능해야 하고 구제되어야 함

○ 유엔 사무총장은 2020년 인권 실현과 신기술의 역할에 대한 보고서를 발표하고, 신기술 도입 관련 의사결정에 이해당사자 참여의 보장과 특히 공공부문의 인공지능 의사결정에 설명가능성 보장 등을 각국 정부에 권고함[12)]

【유엔 사무총장】 사회권의 실현에 있어 신기술의 역할

(a) 신기술의 개발, 사용 및 거버넌스에 있어 모든 인권의 보호 및 강화를 중심 목표로서 전적으로 수용하고, 모든 인권에 대하여 온라인과 오프라인에서 동등한 존중과 이행을 보장해야 한다.

(b) 국가가 민간 부문 활동에 관한 조치를 포함하여 입법 조치를 취해야 할 의무를 재확인하고 준수함으로써, 신기술은 경제・사회・문화적 권리를 포함한 모든 사람들의 인권에 대한 완전한 향유에 기여하고 인권에 미치는 부작용이 방지되어야 한다.

(c) 국가 간 및 국가 내적으로 정보 격차 및 기술 격차를 해소하기 위한 노력을 가속화하고, 신기술의 접근성, 가용성, 경제성, 적응성 및 품질을 개선하기 위한 포괄적인 접근 방식을 촉진해야 한다.

12) Secretary-General (2020). “Question of the realization of economic, social and cultural rights in all countries: the role of new technologies for the realization of economic, social and cultural rights”. 유엔문서번호 A/HRC/43/29(2020. 3. 4). 62문.

(d) 기술 변화 등에 의해 야기되는 변화와 불안정성으로부터 탄력성을 구축할 수 있는 사회적 보호의 권리에 투자하고, 모든 고용 형태의 노동권을 보호해야 한다.

(e) 공공부문에서 신기술, 특히 인공지능의 이용에 관한 정보를 대중에게 전파하기 위한 노력을 대폭 증진해야 한다.

(f) 신기술의 개발 및 도입에 관한 의사결정에 모든 관련 이해당사자의 참여를 보장하고, 특히 공공부문에서 인공지능이 지원하는 의사결정에 대하여 적절한 설명가능성이 보장될 필요가 있다.

(g) 인권의 향유에 중대한 영향을 미칠 수 있는 신기술 시스템, 특히 인공지능 시스템의 전체 생애주기 동안 체계적으로 인권 실사를 실시해야 한다.

(h) 신기술이 사용되는 상황에서 완전한 책임을 보장하는 적절한 법률 체계와 구조를 창출해야 하며, 이는 국내 법제도의 공백을 검토 및 평가하고, 필요한 경우 감독 체제를 수립하고, 신기술로 인한 피해에 대해 접근 가능한 구제 수단을 마련하는 것이 포함된다.

(i) 신기술의 개발 및 사용, 특히 경제・사회・문화적 권리의 향유에 필수적인 제품 및 서비스에 대한 접근에 있어서 차별과 편견을 해소해야 한다.

(j) 정례인권검토(UPR)와 인권조약기구 하에서 이루어지는 보고 및 검토에 있어 신기술이 경제・사회・문화적 권리에 미치는 영향에 특히 주의를 기울여야 한다.

□ 과학기술정보통신부, 인공지능 윤리기준 공개(2020. 12. 23.)[13]

○ 「인공지능 윤리기준」은 ‘사람 중심의 인공지능’을 위한 최고 가치인 ‘인간성(Humanity)’을 위한 3대 기본원칙과 10대 핵심요건을 제시하고 있음

- (목표 및 지향점) ① 모든 사회 구성원이 ② 모든 분야에서 ③ 자율적으로 준수하며 ④ 지속 발전하는 윤리기준을 지향한다.

- (3대 기본원칙) ‘인간성(Humanity)’을 구현하기 위해 인공지능의 개발 및 활용 과정에서 ❶ 인간의 존엄성 원칙, ❷ 사회의 공공선

13) “과기정통부, 사람이 중심이 되는 「인공지능(AI) 윤리기준」 마련”, 과학기술정보통신부 보도자료(2020. 12. 23).

원칙, ❸ 기술의 합목적성 원칙을 지켜야 한다.

- (10대 핵심요건) 3대 기본원칙을 실천하고 이행할 수 있도록 인공지능 개발~활용 전 과정에서 ① 인권 보장, ② 프라이버시 보호, ③ 다양성 존중, ④ 침해금지, ⑤ 공공성, ⑥ 연대성, ⑦ 데이터 관리, ⑧ 책임성, ⑨ 안전성, ⑩ 투명성의 요건이 충족되어야 한다.

【과학기술정보통신부】 인공지능 윤리기준

구분	내용
3대 기본원칙	① 인간 존엄성 원칙 - 인간은 신체와 이성이 있는 생명체로 인공지능을 포함하여 인간을 위해 개발된 기계제품과는 교환 불가능한 가치가 있다. - 인공지능은 인간의 생명은 물론 정신적 및 신체적 건강에 해가 되지 않는 범위에서 개발 및 활용되어야 한다. - 인공지능 개발 및 활용은 안전성과 견고성을 갖추어 인간에게 해가 되지 않도록 해야 한다. ② 사회의 공공선 원칙 - 공동체로서 사회는 가능한 한 많은 사람의 안녕과 행복이라는 가치를 추구한다. - 인공지능은 지능정보사회에서 소외되기 쉬운 사회적 약자와 취약 계층의 접근성을 보장하도록 개발 및 활용되어야 한다. - 공익 증진을 위한 인공지능 개발 및 활용은 사회적, 국가적, 나아가 글로벌 관점에서 인류의 보편적 복지를 향상시킬 수 있어야 한다. ③ 기술의 합목적성 원칙 - 인공지능 기술은 인류의 삶에 필요한 도구라는 목적과 의도에 부합되게 개발 및 활용되어야 하며 그 과정도 윤리적이어야 한다. - 인류의 삶과 번영을 위한 인공지능 개발 및 활용을 장려하여 진흥해야 한다.
10대 핵심요건	① 인권보장 - 인공지능의 개발과 활용은 모든 인간에게 동등하게 부여된 권리를 존중하고, 다양한 민주적 가치와 국제 인권법 등에

	명시된 권리를 보장하여야 한다. - 인공지능의 개발과 활용은 인간의 권리와 자유를 침해해서는 안 된다. ② 프라이버시 보호 - 인공지능을 개발하고 활용하는 전 과정에서 개인의 프라이버시를 보호해야 한다. - 인공지능 전 생애주기에 걸쳐 개인 정보의 오용을 최소화하도록 노력해야 한다. ③ 다양성 존중 - 인공지능 개발 및 활용 전 단계에서 사용자의 다양성과 대표성을 반영해야 하며, 성별・연령・장애・지역・인종・종교・국가 등 개인 특성에 따른 편향과 차별을 최소화하고, 상용화된 인공지능은 모든 사람에게 공정하게 적용되어야 한다. - 사회적 약자 및 취약 계층의 인공지능 기술 및 서비스에 대한 접근성을 보장하고, 인공지능이 주는 혜택은 특정 집단이 아닌 모든 사람에게 골고루 분배되도록 노력해야 한다. ④ 침해금지 - 인공지능을 인간에게 직간접적인 해를 입히는 목적으로 활용해서는 안 된다. - 인공지능이 야기할 수 있는 위험과 부정적 결과에 대응 방안을 마련하도록 노력해야 한다. ⑤ 공공성 - 인공지능은 개인적 행복 추구 뿐만 아니라 사회적 공공성 증진과 인류의 공동 이익을 위해 활용해야 한다. - 인공지능은 긍정적 사회변화를 이끄는 방향으로 활용되어야 한다. - 인공지능의 순기능을 극대화하고 역기능을 최소화하기 위한 교육을 다방면으로 시행하여야 한다. ⑥ 연대성 - 다양한 집단 간의 관계 연대성을 유지하고, 미래세대를 충분히 배려하여 인공지능을 활용해야 한다. - 인공지능 전 주기에 걸쳐 다양한 주체들의 공정한 참여 기회를 보장하여야 한다. - 윤리적 인공지능의 개발 및 활용에 국제사회가 협력하도록

	노력해야 한다. ⑦ 데이터 관리 - 개인정보 등 각각의 데이터를 그 목적에 부합하도록 활용하고, 목적 외 용도로 활용하지 않아야 한다. - 데이터 수집과 활용의 전 과정에서 데이터 편향성이 최소화되도록 데이터 품질과 위험을 관리해야 한다. ⑧ 책임성 - 인공지능 개발 및 활용과정에서 책임주체를 설정함으로써 발생할 수 있는 피해를 최소화하도록 노력해야 한다. - 인공지능 설계 및 개발자, 서비스 제공자, 사용자 간의 책임소재를 명확히 해야 한다. ⑨ 안전성 - 인공지능 개발 및 활용 전 과정에 걸쳐 잠재적 위험을 방지하고 안전을 보장할 수 있도록 노력해야 한다. - 인공지능 활용 과정에서 명백한 오류 또는 침해가 발생할 때 사용자가 그 작동을 제어할 수 있는 기능을 갖추도록 노력해야 한다. ⑩ 투명성 - 사회적 신뢰 형성을 위해 타 원칙과의 상충관계를 고려하여 인공지능 활용 상황에 적합한 수준의 투명성과 설명 가능성을 높이려는 노력을 기울여야 한다. - 인공지능기반 제품이나 서비스를 제공할 때 인공지능의 활용 내용과 활용 과정에서 발생할 수 있는 위험 등의 유의사항을 사전에 고지해야 한다.

○ 국회 입법조사처는 인공지능의 윤리적 사용을 위해서 정부는 인공지능의 윤리기준 등과 관련된 문제를 조정하고 해결할 수 있는 거버넌스나 사후 감시·감독시스템 등을 도입하고, 기업은 △윤리전문가 채용 △인공지능 윤리강령 제정 △인공지능 피해보상 방안 등을 마련하며, 소비자 보호를 위해 인공지능으로 인한 피해에 대한 배상책임 제도를 보완할 필요가 있다고 지적함[14].

14) 이순기 (2020). “인공지능의 윤리적 사용을 위한 개선과제”. 국회입법조사처 이슈와논점 제1759호(2020. 9. 25).

- 특히 입법조사처는 인공지능 챗봇 ‘이루다’ 논란에 대한 후속보고서에서 해외 정책을 참고하여 인공지능 알고리즘의 위험성에 따른 차등적 관리 체계의 검토를 권고하였음[15]. 즉, 고위험 인공지능 알고리즘 분야에서는 엄격한 사전 관리를, 저위험 분야에서는 완화된 관리로 구분하여 규율할 것을 제안함

> 국회 입법조사처 <이슈와 논점> 제1799호 중
>
> 인공지능이 견고함과 신뢰성을 갖출 때 기술에 대한 사람들의 수용성이 높아져 기술발전과 산업 활용이 촉진될 수 있다. 이를 위해 현재 다소 추상적이고 선언적인 인공지능 윤리기준을 보다 구체화하고 검증 가능한 형태로 발전시킬 필요가 있다 … (중략) … 미국・유럽 등의 입법・정책을 참고하여 고위험 분야에서는 사전 점검 체계를, 그 외의 분야에서는 자율 규제 또는 품질 인증 체계 도입을 검토해 볼 수 있을 것이다. 사전 점검의 방안으로는 학습데이터 관리, 투명한 정보 제공, 인간의 개입 등 실효성과 집행가능성 있는 기준들을 마련할 필요가 있다.

ㅇ 과학기술정보통신부는 2022년 하반기를 목표로 “고위험 분야 인공지능 기술 기준 마련”의 계획을 공표함[16]

15) 심용우, 정준화 (2021). 앞의 글. 4p 참조.
16) “디지털 뉴딜 성과 창출 가속화를 위한 법·제도 정비 본격 착수”, 과학기술정보통신부 보도자료(2021. 3. 19). 참고2.

2. 공공기관 인공지능 윤리

□ 해외 규범

○ 영국 국가 전문연구기관인 앨런튜링 연구소는 공공기관 인공지능 시스템이 고려해야 할 해악 우려를 다음과 같이 분류함[17]

【영국 앨런튜링 연구소】 공공부문 인공지능 시스템의 해악 우려

- ▸편향 및 차별
- ▸개인 자율성, 권리구제, 권리행사 거부
- ▸불투명성, 설명불가능성, 부당한 결과
- ▸프라이버시 침해
- ▸사회적 관계 단절 및 고립
- ▸신뢰할 수 없고, 안전하지 않으며, 품질이 낮은 결과물

○ 이후 영국 정부는 2019년 6월 발간한 <공공부문 인공지능 활용 가이드>에서 공공기관이 고려해야 할 위험성과 완화 방안을 다음과 같이 설명함[18]

【영국 정부】 공공부문 인공지능 프로젝트의 위험과 완화 방법

위험	완화 방법
편향 또는 차별의 징후	▸모델의 편향된 결과를 모니터링 하거나 공정하고 설명할 수 있게 만드는 프로세스가 있는지 확인
데이터 사용이 법·제도, 정부기관의 규정을 준수하지 않음	▸인공지능 데이터 준비에 대한 지침을 참조
기밀 유지 및 데이터 무결성 유지를 보장하는 보안 프로토콜이 존재하지 않음	▸필요한 보안 프로토콜을 정의하기 위해 데이터 카탈로그를 구축

17) The Alan Turing Institute (2019). “A guide for the responsible design and implementation of AI systems in the public sector”. p4.
18) Government Digital Service and Office for Artificial Intelligence (2019a). “A guide to using artificial intelligence in the public sector”.

데이터에 접근할 수 없거나 열악한(poor) 데이터 품질	▸내부 및 외부에서 초기 단계에 사용할 데이터셋을 매핑하고, 이후에 데이터를 정확성, 완전성, 고유성, 관련성, 충분성, 적시성, 대표성, 타당성 또는 일관성의 조합에 대한 기준으로 평가
모델 통합 불가능	▸인공지능 모델 구축 초기에 엔지니어를 포함해 개발된 모든 코드가 운용 준비가 되었는지 확인
모델에 대한 책임 프레임워크가 없음	▸인공지능 모델의 서로 다른 영역에서 최종 책임을 지는 책임자를 정의하기 위해 명확한 책임 기록 확립

*요약번역: 한국정보화진흥원(NIA)

- 나아가 <공공부문 인공지능 활용 가이드>는 공공기관이 인공지능을 활용할 때 6가지 요소를 반드시 고려하도록 함[19)]

【영국 정부】 공공기관 인공지능 활용의 고려 요인

구성	세부내용
데이터 품질	▸활용의 성공 여부는 데이터 품질의 우수성이 핵심
공정성	▸인공지능 모델은 관련된 훈련과 테스트가 중요하며, 정확하고 일반화 가능한 데이터셋 활용도 중요 ▸인공지능 시스템이 의도된 목적에 부합해 개발될 수 있도록 전문 지식을 보유한 인력이 개발한 것인가?
책무성	▸인공지능 모델의 각 요소를 담당하는 사람과 인공지능 시스템의 설계자 및 구현자의 최종 책임을 묻는 방법을 고려
개인정보 보호	▸유럽연합 개인정보보호법(GDPR) 및 영국데이터보호법(DPA 2018)과 같은 데이터 법・제도의 준수 여부
설명가능성 및 투명성	▸인공지능 모델이 결론에 도달한 방법은 설명가능한가?

19) Government Digital Service and Office for Artificial Intelligence (2019a). 앞의 문서.

비용	▸인공지능 인프라 구축, 실행 및 유지보수, 관련 인력 훈련 및 교육 등 인공지능 도입 비용과 그에 따른 경제적 효과(혜택, 이익)를 비교

*요약번역: 한국정보화진흥원(NIA)

○ 영국 정부 인공지능사무국은 공공부문을 위해 보다 구체적인 <인공지능의 윤리와 안전을 고려한 시스템 설계·구현 가이드>를 발표함. 공공기관은 인공지능 프로젝트 실행 시 책임 있는 데이터 설계와 활용 체계를 지원·보증하고 동기를 부여해야 함 (SUM 원칙: Support, Underwrite, Motivate)[20]

【영국 정부】 공공부문 인공지능 기술 설계·활용의 윤리적 가치 체계와 실행 원칙

구분		내용
가치 체계	존중	▸개인의 존엄성 회복: 자유롭고 정보에 입각한 결정을 내릴 수 있는 능력을 보장, 자율성·자기표현력 등의 권리를 보호
	연결	▸공개적·포괄적 연결: 인공지능 프로젝트 과정의 전 주기에서 다양성과 참여를 활성화, 사회적인 신뢰와 공감, 상호 책임 및 이해의 체계를 강화
	돌봄	▸복지를 위한 돌봄: 인공지능 시스템에 영향을 받는 모든 사람들의 복지와 안전을 증진, 해당 기술의 오용과 남용 위험을 최소화
	보호	▸사회적 가치와 공익 보호: 모든 사람을 동등하게 대우하고 사회적 형평성을 보호, 인공지능 및 디지털 기술을 법에 따라 공정·균등하게 보호
실행 원칙	공정성	▸데이터 공정성: 공정한 데이터셋을 사용 ▸설계 공정성: 모델에 합리적인 기능, 프로세스 및 분석 구조를 포함 ▸산출 공정성: 결과물이 차별적 영향을 미치지 않도록 함 ▸시행 공정성: 편파적이지 않은 방법으로 제도를 시행

20) Government Digital Service and Office for Artificial Intelligence (2019b). “Guidance: Understanding artificial intelligence ethics and safety”

	책임성	▸프로젝트의 설계 및 구현의 전 과정에 관련된 모든 역할에 책임을 설정 ▸프로젝트 전체 단계에서 검토 및 감독 등의 활동 모니터링 실행
	지속 가능성	▸정확성、신뢰성、보안성、견고성을 포함하여 궁극적으로 안정성을 고려 ▸인공지능 설계자와 사용자는 인공지능 시스템이 개인、사회에 미칠 수 있는 영향 등을 인지해야 함
	투명성	▸인공지능 모델이 처리된 방법과 근거 등을 영향을 받는 이해당사자들에게 공개

*요약번역: 한국인터넷진흥원(KISA). 일부수정.

○ 영국 공직생활윤리위원회는 2020년 보고서에서 향후 공공기관 인공지능 윤리를 실현하기 위한 제도 마련을 다음과 같이 제안함[21)]

- 공공기관이 인공지능 도입에 있어 현행 법률을 준수할 것을 요구

- 또한 정부에 인공지능 공공조달 규칙을 마련하고, 공공기관 인공지능 영향평가의 의무적 실시 및 공개 제도를 마련할 것을 요구함. 더불어 공공기관 인공지능에 대한 정보 공개 기준을 마련할 것 또한 요구하는 한편, 공공부문 인공지능의 평등법 준수지침을 개발할 것을 영국 평등인권위원회에 요구함

【영국 공직생활윤리위원회】 인공지능과 공공 윤리

▸정부/국가기관/규제기관, 공공 및 민간 공공서비스 제공자에 대한 권고

1 윤리적 원칙과 지침 마련
- 정부는 공공부문 인공지능 활용에 대한 세 가지 윤리 원칙(FAST SUM 원칙, OECD 인공지능 원칙, 데이터 윤리 프레임워크)의 목적, 적용범위 및 위상을 명확하게 알려야 함

2 인공지능의 법적 근거 명확화
- 모든 공공 부문 조직은 공공서비스에 대한 인공지능 기술 적용이 관련 법률 및 규정을 준수하는지를 발표해야 함

21) The Committee on Standards in Public Life (2020). “Artificial Intelligence and Public Standards”.

3 데이터 편향 및 차별 금지 지침 마련
- 평등인권위원회는 앨런튜링 연구소 및 데이터윤리혁신센터(CDEI)와 협력하여 공공기관이 평등법을 준수하도록 지침을 개발해야 함

4 규제 보증기구 설립
- 공공영역 인공지능 사용에 대한 규제 보증기구가 있어야 하며, 위원회는 CDEI가 이 역할을 수행하는 것을 지지함

5 조달 규칙 및 절차 윤리기준 마련
- 정부는 공공부문 인공지능 솔루션을 개발하는 민간기업이 공공기준을 충족하도록 조달 요건을 마련해야 하며, 입찰 및 계약 시 윤리기준에 대한 요건을 명시해야 함

6 국영상업서비스의 디지털 시장에서 윤리기준 마련
- 국영상업서비스의 경우에는 인공지능 상품 및 서비스가 공공표준을 준수하는지 또는 훼손하는지에 대해 평가할 수 있는 다양한 기준을 도입하고, 서비스 제공자는 윤리적 요건에 맞는 인공지능 상품 및 서비스를 찾아야 함

7 의무적 영향평가 및 공개
- 정부는 인공지능이 공공표준에 미치는 잠재적 영향에 대한 평가를 현행 절차에 통합하는 방안을 검토해야 함. 이러한 평가는 의무적으로 실시하고 공개되어야 함

8 투명성 및 공개성
- 정부는 공공기관의 인공지능 시스템 관련 신고 및 정보공개에 관한 명확한 지침을 마련해야 함

▸공공서비스를 제공하는 공공 및 민간업체에 대한 권고

9 공공표준에 대한 위험성 평가
- 공공서비스 제공자는 인공지능 시스템이 공공표준에 미칠 잠재적 영향을 평가하고, 시스템 설계가 공공표준에 미칠 위험을 완화하는지 확인해야 함. 인공지능 시스템 설계를 변경할 때마다 표준에 대한 검토가 이루어져야 함

10 다양성 고려
- 공공서비스 제공자는 인구의 다양한 배경, 행동, 관점이 고려되었는지 확인함으로써 편견과 차별이 없는 서비스를 제공하기 위해 노력해야 함

11 책임소재 명확화
- 공공서비스 제공자는 인공지능 시스템에 대한 책임소재를 명확히 해야 함. 인공지능 시스템에 대한 책임을 명확하게 할당하고 문서화해야 하며, 인공지능 시스템 운영자는 책임을 다해야 함

⑫ 모니터링 및 평가
- 공공서비스 제공자는 인공지능 시스템이 원래의 목적에 맞게 운영되고 있는지 항상 모니터링하고 평가해야 함

⑬ 감독 매커니즘 확립
- 공공서비스 제공자는 인공지능 시스템을 적절히 감시할 수 있는 감독 매커니즘을 확립해야 함

⑭ 이의제기 및 배상방법 안내
- 공공서비스 제공자는 시민에게 그들의 권리와 인공지능 기반 결정에 대해 이의제기하는 방법을 알려야 함

⑮ 직원 훈련 및 교육
- 공공서비스 제공자는 인공지능 시스템을 활용하는 직원이 지속적인 훈련 및 교육을 받도록 해야 함

*요약번역: 정보통신정책연구원(KISDI), 일부수정.

○ 한편 세계 여러 나라에서 공공기관 인공지능 의사결정의 편향성 및 차별적 결과에 따른 논란이 불거짐

- 유럽연합은 <인공지능 백서>에서 "인간의 의사결정에도 편견이 작용하지만 인공지능 의사결정에서 작용하는 편견은 통제 메커니즘 없이 훨씬 더 많은 사람들에게 장기간 영향을 준다"고 지적함

【사례】 형사사법 분야 인공지능 의사결정의 차별 위험성[22)]

▸미국 위스콘신주 대법원은 2016년 피고인의 재범 위험성을 평가할 때 참고하는 콤파스(COMPAS) 알고리즘의 평가지수가 법원 결정의 유일한 요소가 되었다면 위법이지만, 보조적인 수단으로 사용되는 경우 적법절차 위반이 아니라고 판결함

▸그러나 언론사 프로퍼블리카에서 2013년부터 2014년까지 콤파스 알고리즘에 의해 법원의 결정이 이루어진 피고인 1200명의 기록을 검증한 결과, 재범률이 높은 것으로 예측되었지만 실제로 2년간 범죄를 저지르지 않은 경우가 흑인의 경우 45%, 백인의 경우는 23.5%이었던 반면, 재범률이 낮은 것으로 예측되었지만 실제로 2년간 범죄를 저지른 경우가 백인이 48%로 흑인 28%보다 훨씬 높았던 것으로 드러남

22) 프로퍼블리카 관련 보도
<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

□ 한국정보화진흥원(NIA), <공공기관 신뢰가능 인공지능 구현 실행가이드> 발표(2019. 12.)[23]

○ ‘OECD 인공지능 권고안’을 적용하여 마련함

【한국정보화진흥원】 공공기관 신뢰가능 인공지능의 구현 실행가이드

원칙	실행가이드
포용성장, 지속가능 발전, 복지증진	공공성의 확인 - 인공지능 시스템의 기관 미션 연계성과 사회경제적 영향평가
	사회적 차별요소 배제 - 데이터, 모델로부터 성, 인종 등 차이로 인한 근원적 차별 배제
인간중심 공정성	인간중심 가치와 공정성 촉진 - 인권영향평가, 인권실사, 윤리 행동 강령, 품질인증 조치
	인간중심 가치 내재화 - 적절한 안전장치 (Kill Switch, Human in the loop 등)
투명성 설명가능성	인공지능 시스템에 대한 투명한 정보공개 인공지능에 관한 일반 정보, 개발/훈련/ 운영/활용의 방식에 관한 정보
	인공지능 시스템 결과에 대한 설명 요인, 데이터, 알고리즘 등 의사결정 요인과 전후 맥락 설명
보안 및 안전성	인공지능 시스템의 추적 가능성 보장 - 데이터 세트, 알고리즘, 프로세스 및 의사결정 관련 추적 가능성
	체계적인 위험관리 접근 - 가능한 위험 및 확률, 관리방안
책임성	인공지능 시스템 원칙의 실현 - 라이프사이클에서 발생한 의사결정과 행동 문서화

23) 한국정보화진흥원 (2019a). “공공기관 신뢰가능 AI 구현 실용가이드: OECD 권고안의 적용”. NIA 「DNA플러스 2019」

3. 공공기관 인공지능 조달 규범

○ 공공조달은 민간경제의 매매계약과 달리 공정성, 절차적 투명성, 책임성을 요하는 특성이 있고, 정보공개법상 공공기관은 관련 자료를 일정부분 공개할 의무도 존재함

- 따라서 조달을 통해 공공부문에 인공지능이 활용될 경우, 관련 절차, 활용 내용, 결과물 등에 대하여 국민은 정보공개 청구를 할 수 있고, 공공기관은 그에 대한 응답하는 과정에서 인공지능의 윤리적 보급과 국민의 신뢰를 확보할 수 있음

□ 해외 규범

○ 유럽연합 집행위원회는 <인공지능 공공조달 백서>에서 “데이터 윤리, 민주주의 및 기본권에 부합하는 공공조달을 구현”하고, 특히 위험성에 따른 체계적 규율을 추진함

- ‘신뢰가능 인공지능’은 “책임성, 기술적 안전성, 지속가능성에 대한 요구 뿐 아니라 데이터 윤리 요소를 포함한 공공조달 체계를 수립하고 이를 현행 법적 의무에 적용함으로써 달성될 수 있다”고 지적함

【유럽연합】 공공조달에 있어 위험기반·체계적 접근법

▸신뢰가능 인공지능은 책임성, 기술적 안전성, 지속가능성에 대한 요구 뿐 아니라 데이터 윤리 요소를 포함한 공공조달 체계를 수립하고 이를 현행 법적 의무에 적용함으로써 달성될 수 있음

▸이를 위한 5단계 실사 절차를 권장함

1 사전적인 위험 영향평가 : 사람과 집단, 권리와 자유, 민주적 조직과 절차, 사회와 환경에 부작용을 미치는지 살핌

2 공급자 예비 심사 : 설계 절차의 최초 단계서부터 다음과 같은 인공지능 관련 데이터 윤리 요건을 고려하고 정의하고 구현해야 함

- 인공지능이 이용자와 직접적으로(챗봇, 가상비서 등) 또는 간접적으로(자동화된 의사결정) 상호작용한다면 이는 필히 인간이 아니라는 점을 밝혀야 함

- 인공지능 시스템이 추적가능하고, 설명가능하고 이해관계자를 수용해야 함
- 인공지능 시스템이 편향을 방지하고 보편적 설계를 따라야 하며 검토 절차를 포함해야 함
- 기술적 안전성은 문서화되어 설명가능성, 공정 커뮤니케이션 및 감사를 보장해야 함

3 계약 : 입찰 선정의 품질 기준은 정보보안, 데이터 윤리, 환경 측면, 프라이버시, 보편적 설계 등에 적용되는 표준 및 관리시스템에 관한 기술사양을 반영해야 함

4 계약 이행 조건 : 발주 공공기관은 계약이행조건에 지속가능성, 기본권 존중, 데이터 윤리에 대한 조항을 포함하고 제재 조항 및 문서화 요건을 명시해야 함

5 계약 집행 : 공급자는 데이터 윤리, 법적 준수, 책임, 기술적 안전성 및 지속가능성의 다섯 가지 표제에 따라 공공 계약에 명시된 요건을 충족해야 함

ㅇ 유럽평의회 인권위원장은 2019년 5월 인공지능에 대한 인권 규제 준수에 대한 보고서에서 공공기관 인공지능 조달 절차에 인권영향평가의 실시를 권고함[24)]

- 공공기관은 인권영향평가의 공표나 수행이 가능하지 않는 공급자로부터 인공지능 시스템을 조달해서는 안 됨

ㅇ 영국 정부는 2020년 6월 인공지능 조달지침을 발표하고 공공조달을 통하는 인공지능에서 10대 원칙을 따르도록 함[25)]

- 특히 이 지침은 조달되는 인공지능 시스템의 “‘블랙박스’ 및 공급업체에 대한 종속(lock-in) 방지”를 요구함

【영국 정부】 인공지능 조달지침

1 인공지능 도입 계획에 조달을 포함할 것
2 다학제적 팀을 구성하여 의사결정을 수행할 것. 낙찰 공급업체에

24) Council of Europe Commissioner for Human Rights (2019). “Unboxing artificial intelligence: 10 steps to protect human rights”.
25) Office for Artificial Intelligence (2020). “Guidelines for AI procurement”.

대해서도, 적합한 기술력을 갖춘 팀을 구성하고 인공지능 시스템의 편향성을 완화하기 위해 다양성 수용을 요구할 것

3 조달절차 개시 전에 데이터 평가를 실시할 것. △조달 절차의 개시 단계부터 데이터 거버넌스 메커니즘이 가동될 수 있도록 확보할 것 △프로젝트에 관련 데이터를 사용할 수 있는지 여부를 평가할 것 △시장에 출시하기 전에 데이터 내부의 결함 및 편향 가능성을 해결할 것. 데이터 문제를 직접 해결할 수 없는 경우 이를 해결하기 위한 계획을 수립할 것 △조달 계획 및 후속 프로젝트를 위해 공급업체와 데이터를 공유할 것인지 여부 및 방법을 정의할 것

4 인공지능 도입의 혜택과 위험성을 평가할 것. △제안서를 평가할 때 공익이 의사결정 절차의 주요 동인이라는 점을 조달 문서에 설명할 것. <사회적 가치 지침>에 따라 인공지능 시스템이 인간과 사회 경제에 미치는 영향 및 편익을 고려할 것. 조달되고 있는 사업이 공익적 목표와 관련이 있어야 하며, 차별금지, 동등한 대우 및 비례성의 원칙을 준수해야 함 △당면한 문제와 관련하여 인공지능을 고려한 배경을 조달 문서에 명확히 설명하고 대안적 솔루션에 대하여 열린 태도를 취할 것 △조달 절차 개시 단계에서 인공지능 영향평가를 수행하고, 중간 조사 결과가 조달에 반영되는지 확인할 것. 주요 의사결정 단계에서 평가 결과를 재차 살펴볼 것

5 시장 형성 초기단계부터 효과적으로 개입할 것. 다양한 인공지능 공급자들과 다양한 방식으로 관계를 맺고, 인공지능 생태계에 개방적인 경쟁 환경을 구축할 것

6 올바른 시장 경로를 구축하고 특정 솔루션보다 해결하고자 하는 과제를 제시할 것

7 거버넌스 및 정보인증을 위한 계획을 수립할 것. 인공지능 시스템에 대한 철저한 검사를 위한 관리감독 메커니즘을 구축하고, 기존 법과 표준을 준수할 것. 인공지능 의사결정의 투명성을 최대화하여 사용자에게 인공지능 시스템이 잘 기능한다는 확신을 부여할 것

8 블랙박스 알고리즘 및 공급업체 종속을 방지할 것. 알고리즘의 설명/해석 가능성을 중요 기준으로 설정하고, 특정 공급업체에 고착되지 않도록 여러 다른 공급자들의 인공지능 시스템 참여를 유도할 것

9 평가 단계에서 인공지능 도입의 기술적/윤리적 한계를 해소 필요성에 집중할 것. 데이터 편향 문제는 없는지, 기존 서비스/기술과 통합 과정에 충분한 검토가 이루어졌는지, 적절한 기술적 표준을 준수하고 있는지 등

10 인공지능 시스템의 생애주기 관리를 고려할 것. △인공지능 조달 과정에서 일회성이 아니라 생애주기에 걸친 검사 필요성을 고려할 것

△지식 이전 및 교육훈련을 요구사항에 포함할 것 △인공지능 시스템을 이해해야 하는 비전문가 대상 교육훈련 및 설명을 요구사항에 포함할 것 △적절하고 지속적인 고객지원 및 호스팅 협의를 보장할 것

*요약번역: 한국과학기술기획평가원(KISTEP), 일부수정.

□ 국내 규범

○ 한국정보화진흥원(NIA)은 2019년 12월 <공공부문 AI 시스템 도입에 따른 조달 분야 이슈 분석> 보고서를 발표함[26)]

- 해외 조달 규범과 비교하여 보았을 때 인공지능 윤리를 소극적으로 반영한 측면이 있음

[한국정보화진흥원] AI 조달 단계별 주요 이슈

단계별 주요 이슈	세부 내용
1 AI 시스템 발주 전(前) 단계의 고려 사항	제안요청서를 작성하기 전, 명확한 목표와 해결하고자 하는 문제점을 명확히 인지하고 현재 보유 데이터의 상태 확인 필수
2 AI 시스템 제안사 선정 을 위한 평가체계 개선	스타트업, 중소기업이 활성화될 수 있는 방안을 모색하고, AI의 특수성을 고려한 기술적 측면 및 윤리적 측면의 평가 필요
3 AI 시스템 구축・운영・ 유지보수 단계의 주요이슈	AI 시스템 운영을 관리할 전문 인력이 필요하고, AI 구축기업과 유지보수 기업 간 발생할 수 있는 이슈 해결 필요

○ 한편, 현행 소프트웨어산업 진흥법령에 따라 공공 발주처가 소프트웨어사업을 추진하는 경우 소프트웨어사업 영향평가를 할 의무가 있음

- 소프트웨어사업영향평가(소프트웨어산업 진흥법 제14조의2 및 동법 시행령 제12조의2)는 “민간 소프트웨어와의 유사성, 민간시장 침해 가능성 및 사업의 필요성・공공성” 등을 종합적으로 평가한 후 결

26) 한국정보화진흥원 (2019). “공공부문 AI 시스템 도입에 따른 조달 분야 이슈 분석”. NIA, 「IT & Future Strategy 보고서」(2019. 12. 31.)

과서를 작성하여야 하고, 만약 추진이 부적합하다고 판단하는 경우 사업내용을 조정하거나 사업을 재검토 하여야 함

4. 기타 법령

□ 개인정보 보호법

○ 개인정보 보호위원회는 지방자치단체 인공지능 단속 시스템에 대한 결정에서 개인정보를 수집·이용할 때 현행 법률 조항의 적법한 준수를 요구함

【사례】 인공지능 수사와 개인정보보호법

서울특별시 민생사법경찰단이 범죄 수사와 내사의 전단계에서 이루어지는 조사활동을 목적으로 인공지능 기반 시스템을 개발하여 인터넷·SNS 등 온라인에 공개된 게시물을 광범위하게 수집하고, 범죄 관련성이 높다고 판단되는 게시물을 분석하여 해당 게시물에 포함된 성명, 아이디, (휴대)전화번호, 주소, 업체명 등을 이용하는 것은 「개인정보 보호법」 제15조 제1항 제1호, 제2호 및 제3호에 위반된다 (개인정보 보호위원회 결정 제2019-09-130호).

○ 개인정보 보호법 및 『개인정보 보호 법령 및 지침·고시 해설서 (2020. 12)』에 따르면, 인공지능 시스템 및 서비스 개발 및 운용을 위하여 개인정보를 수집·이용하려면 다음의 경우에 해당하여야 함

개인정보 보호법

제15조(개인정보의 수집、이용) ① 개인정보처리자는 다음 각 호의 어느 하나에 해당하는 경우에는 개인정보를 수집할 수 있으며 그 수집 목적의 범위에서 이용할 수 있다.

1. 정보주체의 동의를 받은 경우
2. 법률에 특별한 규정이 있거나 법령상 의무를 준수하기 위하여 불가피한 경우
3. 공공기관이 법령 등에서 정하는 소관 업무의 수행을 위하여 불가피한 경우
4. 정보주체와의 계약의 체결 및 이행을 위하여 불가피하게 필요한 경우
5. 정보주체 또는 그 법정대리인이 의사표시를 할 수 없는 상태에 있거나 주소불명 등으로 사전 동의를 받을 수 없는 경우로서 명백히 정보주체

또는 제3자의 급박한 생명, 신체, 재산의 이익을 위하여 필요하다고 인정되는 경우

6. 개인정보처리자의 정당한 이익을 달성하기 위하여 필요한 경우로서 명백하게 정보주체의 권리보다 우선하는 경우. 이 경우 개인정보처리자의 정당한 이익과 상당한 관련이 있고 합리적인 범위를 초과하지 아니하는 경우에 한한다.

- 개인정보 보호법 관련 규정 및 『교육분야 가명정보 처리 가이드라인(2020. 10)』에 따르면, 가명정보는 법령에서 허용하는 "정당한 처리 범위 내에서 통계작성, 과학적 연구, 공익적 기록보존 등의 목적으로 정보주체의 동의 없이" 처리 가능함

- 다만, 챗봇 이루다 사건의 경우 개발사는 다른 서비스에서 수집한 개인정보를 비식별조치(가명처리) 후 문제 없이 이용하였다고 주장하였으나 개인정보 보호법 위반 혐의로 조사를 받고 있다는 점에서 주의를 요함

○ 한편 '얼굴인식' 등 생체인식 정보의 경우 '민감정보'로서 일반 개인정보 보다 특별한 보호를 받고 있으며, 특히 해외에서는 아동학생의 생체인식에 대한 이중적인 보호를 적용하고 있음

【사례】 유럽 각국, 학생 얼굴인식 기술에 개인정보 보호법 위반으로 결정[27)]

▸스웨덴의 한 지방자치단체가 얼굴인식 기술을 사용하여 학생들의 학교 출석을 모니터링한 데 대하여 스웨덴 개인정보 보호위원회는 유럽연합 개인정보 보호법 위반으로 결정하고 벌금을 부과함

▸마르세유와 니스 지역의 고등학교에서 보안상의 이유로 얼굴인식 기술을 사용하는 것에 대하여 프랑스 개인정보 보호위원회는 개인정보 보호법 위반으로 보았고, 이후 프랑스 법원 또한 이를 위법으로 판결함

27) FRA(2019), "Facial recognition technology: fundamental rights considerations in the context of law enforcement". 각주47. ; "FIRST EVER DECISION OF A FRENCH COURT APPLYING GDPR TO FACIAL RECOGNITION" , <https://ai-regulation.com/first-decision-ever-of-a-french-court-applying-gdpr-to-facial-recognition/>

ㅇ 개인정보 보호위원회는 「AI 환경의 개인정보보호 수칙」을 마련 중에 있으며, 검토 중인 주요 원칙 및 실천수칙(예시)은 다음과 같음[28)]

<주요 원칙 및 실천수칙(예시)>

- (적법성) 이용자가 개인정보 수집·이용 목적 등을 명확히 인지하도록 사전동의
- (안전성) 개인정보의 비식별처리 활용 및 암호화, 유·노출 방지 등 안전조치
- (투명성) 개인정보의 활용 범위 및 보유기간, AI 서비스 작동흐름 등 공개

ㅇ 나아가 인공지능 등 '자동화 의사결정에 대한 배제 등의 권리 도입'을 내용으로 하는 개인정보 보호법의 개정이 추진되고 있음(입법예고 2021. 1. 6. ~ 2. 16.)

- 개인정보 보호법 개정안 에서는 자동화 의사결정 등에 대한 거부, 이의제기, 설명요구권 등의 제도 신설을 예정하고 있음(개정안 제37조의2)

□ 차별금지 관련법

ㅇ 우리나라에 비록 포괄적인 차별금지법 또는 평등법은 아직 존재하지 않지만, 개별적 차별금지법이 영역별로 규율하고 있음.

- 현행 법률 중 차별 행위를 시정하는 절차를 규율하고 있는 법률로는 △국가인권위원회법, △장애인차별금지 및 권리구제 등에 관한 법률(장애인차별금지법), △남녀고용평등과 일·가정 양립 지원에 관한 법률(남녀고용평등법), △고용상 연령차별금지 및 고령자고용촉진에 관한 법률(연령차별금지법), △기간제 및 단시간근로자 보호 등에 관한 법률, 그리고 △파견근로자 보호등에 관한 법률(비정규직차별금지법) 등을 들 수 있음. 특히 국가인권위원회법은 19가지 차별사유 및 고용·거래·교육 등의 차별영역을 아우르는 '일반적 차별금지법'

28) "개인정보 보호위원회 토론문". <인공지능의 공정성·투명성·책임성 보장을 위한 법제정비방안 토론회(정필모 의원 등 주최, 2021. 2. 17)>.

으로 분류됨

- 특히 공공기관이 도입 및 사용하는 인공지능 서비스 및 제품에 있어서는 현행 법률에서 금지하고 있는 차별이 발생하지 않도록 훈련데이터의 사전조치 및 결과의 공정성을 유지할 의무가 있다 할 것임

□ 적법절차 의무

○ 적법절차원칙이란, “국가공권력이 국민에 대하여 불이익한 결정을 하기에 앞서 국민은 자신의 견해를 진술할 기회를 가짐으로써 절차의 진행과 그 결과에 영향을 미칠 수 있어야 한다는 법원리”를 말함

- 헌법재판소는 “헌법상 적법절차의 원칙은 형사절차뿐만 아니라 입법과 행정 등 국가의 ‘모든’ 공권력 행사에 적용된다”고 보았음(헌법재판소 1992. 12. 24 선고 92헌가8 결정 등). 대법원 역시 “헌법상 적법절차의 원칙은 형사소송절차 뿐만아니라 국민에게 부담을 주는 행정작용에서도 준수되어야 한다”고 판시한 바 있음(대법원 2012. 10. 18 선고 2010두1234 판결).

- 특히 행정기관의 적법절차 의무는 “행정절차와 행정적 결정에 있어서 불공평한 대우, 불공정한 대우 그리고 적절한 기간을 넘어서는 행정절차적 진행”을 제한함

- 따라서 행정기관을 비롯한 공공기관이 인공지능을 이용하여 자동화된 의사결정에 이를 때에는 청문권, 문서열람권, 결정의 이유제시요구권, 이의신청 및 권리구제 등 적법절차를 보장할 필요가 있음

Ⅲ. 해외 인공지능 알고리즘 등급 관련 규범 검토

1. 캐나다 정부 <자동화된 의사결정에 대한 지침>

○ 2019년 캐나다 정부는 <자동화된 의사결정 지침(재정위원회 훈령)>을 발표하여 공공기관 인공지능 요건을 법규화함[29)]

【캐나다 정부】 자동화된 의사결정에 대한 지침 (요건)

6. 요건

자동 결정 시스템 사용 프로그램을 주무하는 부처의 실장이 지명하는 사람 또는 차관보는 다음을 책임진다.

6.1. 알고리즘 영향평가

6.1.1. 자동화된 의사결정 시스템을 생산하기 전에 알고리즘 영향평가를 완료한다.

6.1.2. 알고리즘 영향평가에 의해 결정이 내려진 경우 부록 C에 규정된 관련 요건을 적용한다.

6.1.3. 자동화된 의사결정 시스템의 기능 또는 범위가 변경될 시 알고리즘 영향평가를 갱신한다.

6.1.4. 알고리즘 영향평가의 최종 결과를 <정부 개방 지침>에 부합하도록 캐나다 정부 웹사이트 및 캐나다 재정위원회가 지정한 기타 서비스를 통해 일반 접근이 가능한 형식으로 공개한다.

6.2. 투명성

의사결정 전 공지

6.2.1. 해당 의사결정이 부록 C에 규정된 바대로 자동화된 의사결정 시스템에 의해 전체 또는 부분적으로 수행된다는 내용을 관련 웹사이트에 공지한다.

6.2.2. <Canada.ca 컨텐츠 스타일 가이드>에 부합하는 뚜렷하고 쉬운 용어를 이용하여 공지한다.

29) The Government of Canada. Directive on Automated Decision-Making. <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592>

의사결정 후 설명

6.2.3. 부록 C에 규정된 대로 결정이 내려진 방법과 이유에 대해 영향을 받는 개인들에게 이해가능하게 설명한다.

구성 요소에 대한 접근 권한

6.2.4. 소프트웨어 구성요소에 대하여 <정보 기술 관리 지침>의 C.2.3.8장에 명시된 요건에 따라 적절한 라이선스를 정한다.

6.2.5. 독점 라이선스를 사용하는 경우 다음을 보장한다.

6.2.5.1. 자동화된 의사결정 시스템에 사용되는 독점 소프트웨어 구성요소의 모든 공개 버전을 해당 부서에 전달하고 보호한다.

6.2.5.2. 캐나다 정부는 특별 감사, 조사, 검사, 심사, 집행 조치 또는 사법 절차에 필요한 경우 자동화된 의사결정 시스템에 대하여 접근하고 시험할 권리를 보유한다. 이는 독점 소프트웨어의 모든 공개 버전에도 적용되며, 이때 인가되지 않은 공개에 적용되는 안전조치를 준수한다.

6.2.5.3. 이러한 접근권의 일부로서, 캐나다 정부는 필요한 경우 외부인에게 이러한 구성요소를 검토하고 감사할 수 있는 권한을 부여할 수 있다.

소스 코드 공개

6.2.6. 다음과 같은 경우를 제외하고, <정보 기술 관리 지침> C.2.3.8장에 명시된 요건에 따라 캐나다 정부가 소유한 사용자 정의 소스 코드를 공개한다.

6.2.6.1. 소스 코드가 1급 비밀, 2급 비밀, C급 대외비로 분류된 데이터를 처리하는 경우

6.2.6.2. 정보공개법에 따라 공개가 면제되거나 제외되는 경우

6.2.6.3. 캐나다 정보관리 최고책임자에 의해 면제되는 경우

6.2.7. 공개된 소스 코드에 대하여 적절한 접근 제한을 결정한다.

6.3. 품질보증

검사 및 모니터링 결과

6.3.1. 생산에 착수하기 전, 자동화된 의사결정 시스템이 사용하는 데이터와 정보에 대하여 의도하지 않은 데이터 편향 및 결과에 부당하게 영향을 미칠 수 있는 기타 요소에 대해 검사할 수 있는 절차를 개발한다.

6.3.2. 자동화된 의사결정 시스템을 의도하지 않은 결과로부터 보호하고

본 지침뿐만 아니라 기관 및 프로그램 관련 법률의 준수를 확인하기 위하여 그 결과를 정기적으로 모니터링하는 절차를 개발한다.

데이터 품질

6.3.3. 자동화된 의사결정 시스템을 위해 수집되고 사용되는 데이터가 정보관리 정책 및 개인정보보호법에 따라 관련성이 있고, 정확하며, 최신인지 검증한다.

전문가 검토

6.3.4. 부록 C에 규정된 바대로 자동화된 의사결정 시스템을 검토하기 위해 적절한 자격을 갖춘 전문가의 자문을 받는다.

직원의 교육훈련

6.3.5. 부록 C에 규정된 바대로 자동화된 의사결정 시스템의 설계, 기능 및 구현에 대한 적절한 직원의 교육훈련을 실시하여 그 운영을 검토, 설명 및 감독할 수 있도록 한다.

비상 계획

6.3.6. 부록 C에 따라 비상 시스템 또는 절차를 수립한다.

보안

6.3.7. <정부 보안 정책>에 따라 시스템의 개발 주기 동안 위험 평가를 실시하고 적절한 안전조치 적용을 확립한다.

합법성

6.3.8. 자동화된 의사결정 시스템의 사용이 해당 법률 요건을 준수하도록 보장하기 위해 기관 법무부서와 협의한다.

인적 개입 보장

6.3.9. 적절한 경우 부록 C에 따라 자동화된 의사결정 시스템이 인적 개입을 허용하도록 보장한다.

6.3.10. 부록 C에 따라 자동화된 의사결정 시스템을 생산하기 전에 적절한 수준의 승인을 획득한다.

6.4. 상환 청구

6.4.1. 고객이 행정 결정에 이의를 제기할 때 사용할 수 있는 상환 청구 적용 옵션을 제공한다.

6.5. 보고

6.5.1. 프로그램 목표 달성에 있어 자동화된 의사결정 시스템의 효과와 효율성에 관한 정보를 캐나다 재무부가 지정한 웹사이트 또는 서비스에 게시한다.

○ 캐나다 <자동화된 의사결정 지침>은 자동화된 의사결정에 사용하는 인공지능 알고리즘을 도입하는 공공기관이 영향평가를 실시하고 위험성 수준별로 4단계로 나누어 관리하도록 함[30]

수준	세부사항
Ⅰ	의사결정이 다음 사항에 거의 영향을 미치지 않을 것으로 보이는 경우 - 개인 또는 공동체의 권리, - 개인 또는 공동체의 건강 또는 복리, - 개인, 단체 또는 공동체의 경제적 이익 - 생태계의 지속가능성 통상 수준 Ⅰ의 결정은 복구되고 일시적인 영향을 미칠 수 있음
Ⅱ	의사결정이 다음 사항에 중간 정도의 영향을 미칠 것으로 보이는 경우 - 개인 또는 공동체의 권리, - 개인 또는 공동체의 건강 또는 복리, - 개인, 단체 또는 공동체의 경제적 이익 - 생태계의 지속가능성 통상 수준 Ⅱ의 결정은 복구시킬 수 있고 단기적인 영향을 미칠 수 있음
Ⅲ	의사결정이 다음 사항에 큰 영향을 미칠 것으로 보이는 경우 - 개인 또는 공동체의 권리, - 개인 또는 공동체의 건강 또는 복리, - 개인, 단체 또는 공동체의 경제적 이익

30) 앞의 문서 중 (Appendix B) Impact Assessment Levels

	- 생태계의 지속가능성 통상 수준 Ⅲ의 결정은 복구되기 어려울 수 있고 지속적인 영향을 미칠 수 있음
Ⅳ	의사결정이 다음 사항에 매우 큰 영향을 미칠 것으로 보이는 경우 - 개인 또는 공동체의 권리, - 개인 또는 공동체의 건강 또는 복리, - 개인, 단체 또는 공동체의 경제적 이익 - 생태계의 지속가능성 통상 수준 Ⅳ의 결정은 복구가 불가능하고 영구적인 영향을 미칠 수 있음

○ 알고리즘 영향평가 결과에 따라 공공기관 인공지능 알고리즘의 수준이 결정되면 4단계로 전문가 검토, 공지, 인적 개입, 설명, 검사, 모니터링, 교육훈련 비상 계획, 시스템 구동 승인 의무 등 훈령상 요건을 차등 적용하도록 함[31)]

31) 앞의 문서 중 (Appendix C) Impact Level Requirements

【캐나다 정부】 <자동화된 의사결정 지침> 알고리즘 영향 수준별 적용 요건

요건	수준 Ⅰ	수준 Ⅱ	수준 Ⅲ	수준 Ⅳ
전문가 검토 (peer review)	비해당	다음중 1개 이상 수행 - 연방, 주, 준주 또는 시 정부기관에서 자격이 인증된 전문가의 검토 - 고등교육기관 학부 유자격 구성원의 검토 - 관련 비정부기구 소속 유자격 연구자의 검토 - 관련 전문성을 갖춘 서드파티 공급자와 계약 - 자동화된 의사결정 시스템의 사양을 전문가가 검토하는 저널에 게재 - 재정위원회 사무처에서 지정한 데이터 및 자동화 자문기구의 검토		다음중 2개 이상 수행하거나 - 캐나다 국립연구위원회, 캐나다 통계청, 또는 캐나다 통신보안기구에서 자격이 인증된 전문가의 검토 - 고등교육기관 학부 유자격 구성원의 검토 - 관련 비정부기구 소속 유자격 연구자의 검토 - 관련 전문성을 갖춘 서드파티 공급자와 계약 - 재정위원회 사무처에서 지정한 데이터 및 자동화 자문기구의 검토, 또는 - 자동화된 의사결정 시스템의 사양을 전문가가 검토하는 저널에 게재
공지	비해당	프로그램이나 서비스 웹사이트에 쉬운 용어로 된 공지 게시	관련 웹사이트에 자동화된 의사결정 시스템에 대한 쉬운 용어로 된 문서 발간하며 다음 사항을 포함할 것 - 구성 요소의 작동 방식 - 행정 결정을 지원하는 방식 - 모든 검토 또는 감사의 결과 - 훈련 데이터에 대한 설명, 또는 이 데이터를 공개적으로 사용할 수 있는 경우 익명화된 훈련 데이터에 대한 링크	

주요 의사결정에 대한 인적 개입(Human-in-the-loop for decisions)	의사결정이 인간의 직접적인 개입 없이 내려질 수 있음		의사결정 절차에서 특정 시점에 인적 개입이 없으면 의사결정이 내려질 수 없음. 더불어 최종 의사결정은 사람에 의해 이루어져야 함	
설명 요건	해당 법정 요건에 추가적으로 공통적인 의사결정 결과에 대하여 유의미한 설명이 제공되도록 보장할 것. 여기에는 웹사이트 자주 묻는 질문 코너(FAQ)를 통해 설명을 제공하는 것이 포함될 수 있음	해당 법정 요건에 추가적으로 수혜, 서비스, 기타 규제 조치를 거부하는 의사결정 결과에 대하여 요청이 있을 경우 유의미한 설명이 제공되도록 보장할 것	해당 법정 요건에 추가적으로 수혜, 서비스, 기타 규제 조치를 거부하는 모든 의사결정 결과에 대하여 유의미한 설명이 제공되도록 보장할 것	
검사 (testing)	- 생산에 착수하기 전, 훈련 데이터가 의도하지 않은 데이터 편향 및 결과에 부당하게 영향을 미칠 수 있는 기타 요소에 대해 검사할 수 있는 적절한 절차를 개발할 것 - 자동화된 의사결정 시스템에서 사용 중인 데이터가 여전히 관련성이 있고 정확하며 최신인지 확인하기 위해 정기적으로 검사할 것			
모니터링	자동화된 의사결정 시스템의 결과를 지속적으로 모니터링하여 의도하지 않은 결과로부터 보호하고 본 지침뿐만 아니라 기관 및 프로그램 관련 법률의 준수를 보장할 것			
교육훈련	비해당	시스템의 설계 및 기능에 대한 문서화	시스템의 설계 및 기능에 대한 문서화. 교육 과정 이수 필수.	시스템의 설계 및 기능에 대한 문서화. 교육 과정 반복적 이수. 교육 이수 확인 수단 마련.
비상 계획 (Contingency Planning)	비해당		자동화된 의사결정 시스템을 사용할 수 없는 경우 비상 계획 및 백업 시스템을 사용할 수 있도록 보장할 것	
시스템 구동 승인	비해당	비해당	부처 실장 승인	재정위원회 승인

2. 뉴질랜드 정부 <위험성 매트릭스>

ㅇ 뉴질랜드 정부는 2020년 7월 <아오테아로아 뉴질랜드 알고리즘 헌장>을 발표하고 공공기관들이 서약하도록 함[32)]

- 뉴질랜드 교육부, 아동부, 교육평가청 등 26개 공공기관이 서약함 (2020. 11. 현재)

【뉴질랜드 정부】 아오테아로아 뉴질랜드 알고리즘 헌장

본 헌장은 정부 기관이 알고리즘을 사용하는 방법에 대해 뉴질랜드인이 신뢰를 갖게끔 하는 약속이다. 본 헌장은 정부가 데이터 사용에 있어 투명성과 책무성을 보여주는 다양한 방법 중 하나다. 하지만, 마오리 데이터 주권(Māori Data Sovereignty)과 같은 중요한 사항은 복합적이며 별도의 검토가 필요하기 때문에 충분히 다룰 수 없었다.

약속

우리 기관은 알고리즘을 이용해 만들어낸 결정이 뉴질랜드 국민에게 영향을 미친다는 것을 이해한다. 우리는 우리의 알고리즘에 기반한 결정의 영향을 평가할 것을 약속한다. 더 나아가 우리는 확인된 위험성 등급에 따라 알고리즘 헌장 약속을 적용할 것을 약속한다.

알고리즘 헌장은 다음을 약속한다.

투명성

의사결정이 알고리즘에 의해 어떻게 영향을 받았는지 명확하게 설명함으로써 투명성을 유지한다. 이는 다음을 포함한다.

▸ 알고리즘을 평이한 영어로 문서화

▸ 데이터와 처리과정에 관한 정보를 접근가능하게 만들기(법적인 제한이 있지 않는 이상)

▸ 데이터가 어떻게 수집되고, 저장되고, 보호받는지에 관한 정보 공개

파트너쉽

다음과 같은 조약의 약정을 통해 명확한 공익을 제공한다.

32) Algorithm charter for Aotearoa New Zealand. <https://data.govt.nz/use-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter/>

▸와이탕이 조약의 원칙에 부합하는 알고리즘의 사용 및 마오리족 세계관(Te Ao Māori)의 관점을 포함하는 알고리즘 개발

사람

다음과 같은 방법으로 사람에 초점을 맞춘다.

▸알고리즘에 관심이 있는 사람, 집단, 커뮤니티를 찾고 적극적인 참여 조직, 알고리즘 사용에 영향을 받는 사람들의 의견 수렴

데이터

다음과 같은 방법으로 데이터가 그 목적에 부합하는지 확인한다.

▸그 한계를 이해하기

▸편향을 식별하고 관리하기

개인정보 보호, 윤리 그리고 인권

다음과 같은 방법으로 개인정보, 윤리 그리고 인권에 대한 보호를 보장한다.

▸의도치 않은 결과를 평가하고 이에 대한 조치를 취하기 위한 정기적인 전문가 검토

인간의 감독

다음과 같은 방법으로 인간의 감독을 유지한다.

▸알고리즘에 대한 공개적인 조사(public inquiry)를 위한 담당자 지명

▸알고리즘에 영향을 받은 결정에 대해 불만을 접수하거나 이의를 제기하기 위한 수단 제공

▸알고리즘에 영향을 받은 결정에서 인간의 역할에 대한 명확한 설명

○ 뉴질랜드 정부의 알고리즘 헌장의 경우, 각 기관이 도입 및 사용하는 인공지능 알고리즘의 위험성 발생가능성과 영향을 평가하고, 평가 결과에 따른 위험성 등급에 의해 헌장의 적용 여부를 결정하도록 함

【뉴질랜드 정부】 알고리즘 영향평가와 위험성 매트릭스

위험성 발생가능성과 영향을 평가하기

뉴질랜드 <알고리즘 평가 보고서>에서는 고급 분석 및 데이터 사용이 공공 서비스 제공에 필수적이라는 사실을 발견했다. 기관이 모든 업무 규칙과 과정에 헌장을 적용하는 것은 불가능하며 헌장의 편익적 취지도 달성할 수 없을 것이다.

그러나 정부 기관에서 사람들의 복리에 중대한 영향을 미칠 수 있는 방식으로 알고리즘을 도입하거나 많은 사람들이 의도치 않은 부작용을 겪을 가능성이 높은 알고리즘을 도입하는 경우에는 헌장을 적용하는 것이 합당하다.

헌장 서명기관은 아래 위험성 매트릭스를 사용하여 자기관 알고리즘 결정을 평가할 수 있다. 이런 평가는 전반적인 위험성 수준을 도출하기 위하여 상대적인 영향 수준에서 의도치 않은 부작용의 발생가능성을 정량화함으로써 기관의 판단을 지원할 것이다.

위험성 등급에 의해 헌장의 적용 여부가 결정된다.

위험성 매트릭스

발생가능성

통상 있음 표준적인 작동 중에 자주 발생할 수 있음			
때때로 있음 표준적인 작동 중에 간혹 발생할 수 있음			
거의 없음 표준적인 작동 중에 발생할 가능성이 낮지만 발생할 수는 있음			
영향 정도	**낮음** 의사결정의 영향이 독자적이며, 심각하지 않음	**중간** 의사결정의 영향이 중간 규모의 사람들에게 미치며, 어느 정도 심각성이 있음	**높음** 의사결정의 영향이 광범위하며, 매우 심각함

위험성 등급

낮음 알고리즘 헌장이 적용될 수 있음	중간 알고리즘 헌장이 적용됨	높음 알고리즘 헌장이 반드시 적용돼야함

적용 및 책무

본 헌장은 서명기관마다 다르게 적용된다. 위험성 매트릭스 접근방식은 서명기관이 위험성이 높은 의사결정에 우선적으로 주력하도록 하고, 정부기관이 법률상 소관과 통상적인 업무 수행을 위해 매일 사용하는 대부분의 업무 규칙을 제외할 수 있게 한다. 그 취지는 뉴질랜드인에게 의도치 않은 해악을 미칠 위험도가 높거나 심각한 알고리즘의 사용에 초점을 맞추려는 데 있다. 이 책무는 영역 검토의 일부로써 12개월 후에 검토될 예정이다.

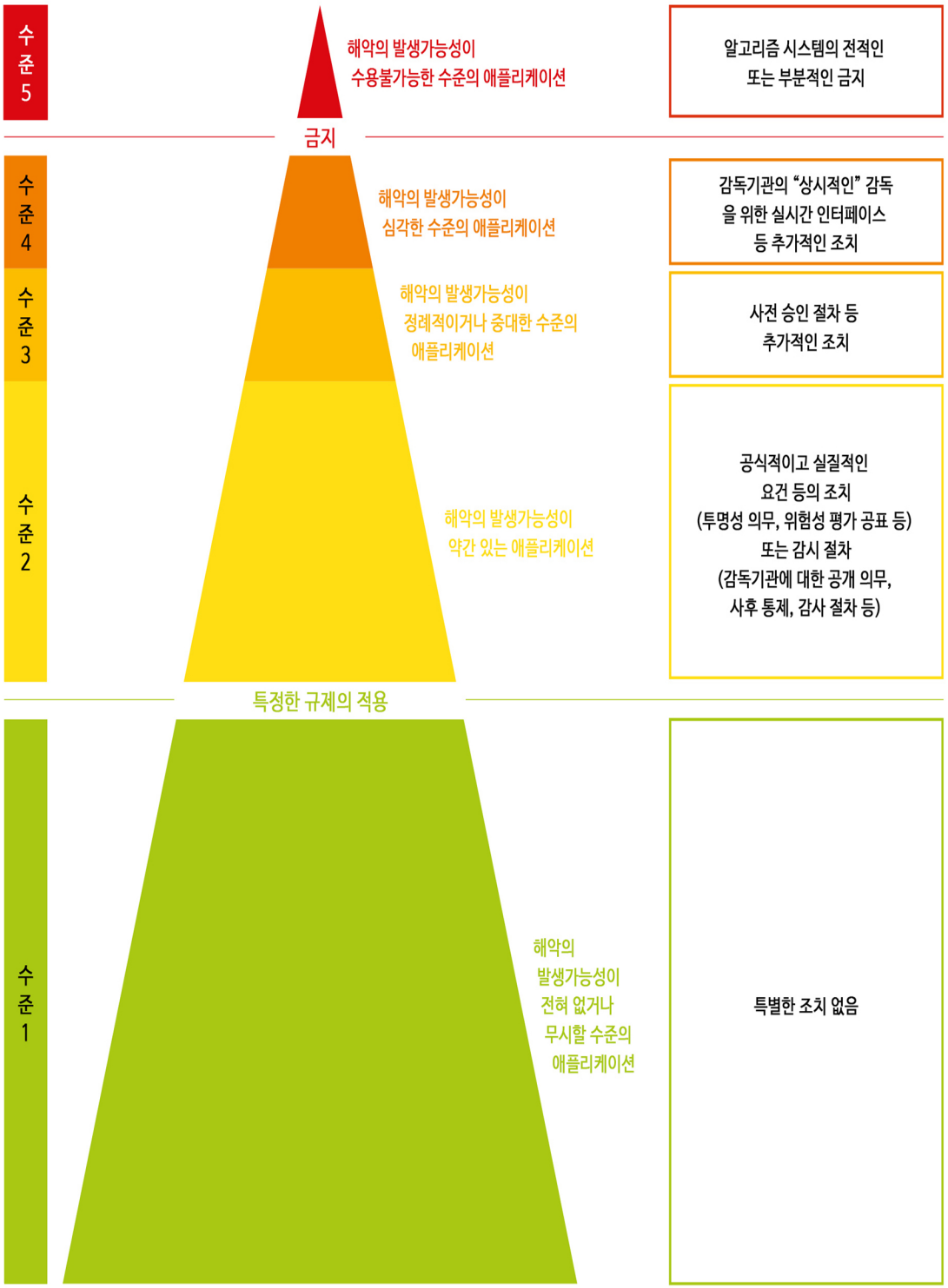
3. 독일 정부 데이터윤리위원회 <알고리즘 시스템 위험도 피라미드>

○ 독일 연방정부 데이터윤리위원회는 2019년 10월 5단계 '위험도 피라미드' 등 위험성에 기반한 알고리즘 시스템의 규제 모델을 제안함[33]

- 2018년 9월 발족한 데이터윤리위원회는 16명의 다학제 전문가들이 참여하는 독립된 위원회로서 연방법무부, 연방내무부가 그 활동을 지원하고 있음

33) Data Ethics Commission (2019). "Opinion of the Data Ethics Commission". <https://assets.contentstack.io/v3/assets/blt3de4d56151f717f2/blt300ce23c9789e0f3/5e5cfe13fa08326331360f93/191023_DEK_Kurzfassung_en_bf.pdf>

【독일 정부】 알고리즘 시스템 위험도 피라미드 및 위험성 기반 규제 시스템



4. 유럽연합 <위험기반 접근법>

ㅇ 2020년 유럽연합 집행위원회는 <인공지능 백서(2020. 2)>[34], <인공지능 공공조달 백서(2020. 5)>[35]를 연달아 발표하며 인공지능 규제 프레임워크를 제시함

- 인공지능이 정부 및 공공기관과 시민 간의 관계를 변화시키고 형성하고 있으며, 공공기관은 유럽이 지향하는 신뢰가능 인공지능의 혁신을 이끌어야 함. 신뢰가능 인공지능은 책임성, 기술적 안전성, 지속가능성에 대한 요구 뿐 아니라 데이터 윤리 요소를 포함한 공공조달 체계를 수립하고 이를 현행 법적 의무에 적용함으로써 달성될 수 있음(인공지능 공공조달 백서)

- 인간의 의사결정에도 편견이 작용하지만 인공지능 의사결정에서 작용하는 편견은 통제 메커니즘 없이 훨씬 더 많은 사람들에게 장기간 영향을 줄 수 있음. 이러한 편견은 인공지능 알고리즘의 설계 모델에서 기인할 수 있고 학습용 데이터의 품질과 훈련 과정에서 기인할 수도 있음(인공지능 백서)

- 유럽 시민들은 기본권(개인정보 보호, 프라이버시, 차별금지 등)과 안전과 책임(소비자 보호, 제품 안전 등)에 관한 법률상 보호수준을 인공지능에 대해서도 기대하고 있음. 고용, 직업, 재화 및 서비스에서 인종, 성별, 장애에 대한 동등한 접근과 처우, 평등에 관한 지침들과 소비자 보호 규정은 물론, GDPR 등 개인정보 보호와 프라이버시 관련 유럽연합의 법률들은 인공지능 여부와 관계없이 원칙적으로 충분히 적용 가능함

- 다만 기존 법률의 효과적인 적용 및 집행, 적용 범위의 불확실성, 행위자 간의 책임 배분 등의 문제를 해결하기 위한 법률 체계를 조정할 필요가 있을 수 있음

ㅇ 유럽연합은 <인공지능 백서>에서 위험기반 인공지능 규제 프레임워크를 제안함. 즉, ‘고위험 인공지능’ 시스템으로 분류된 제품 및

34) European Commission (2020a). 앞의 문서.
35) European Commission (2020b). 앞의 문서.

서비스에 대하여 법적 의무를 부과하고 이에 대한 준수를 검증하기 위한 사전 적합성 검사의 실시를 의무화함

【유럽연합】 고위험 인공지능의 범주

구분	범주
일반	①일반적으로 수행되는 활동의 특성이 상당한 위험이 발생할 것으로 예상되는 분야에서 사용되고(의료, 운송, 에너지 및 공공부문 등) ②해당 인공지능 애플리케이션이 해당 분야에서 상당한 위험이 발생할 가능성이 높은 방법으로 사용되는 경우(개인이나 기업의 권리에 법적인 영향 또는 유사하게 상당한 영향을 미치는 경우. 또는 부상·사망 또는 상당한 물질적·비물질적 손상을 초래하거나, 개인이나 법인이 합리적으로 피할 수 없는 영향을 미치는 경우)
특별	채용 과정과 근로자의 권리에 영향을 미치는 인공지능 애플리케이션의 사용
	소비자 권리에 영향을 미치는 인공지능 애플리케이션의 사용
	원격 생체인식 및 기타 침입 감시 기술 목적의 인공지능 애플리케이션의 사용

- 고위험 인공지능에 대해서 기존 법률 규제에 더한 요구사항이 법적 의무로 적용됨. 이러한 법적 의무로는 △훈련 데이터 △기록과 데이터의 보존 △정보 제공 △견고성 및 정확성 △인적 감독 △원격 생체인식 등 특정 애플리케이션 특별 요건 등이 제시됨

【유럽연합】 고위험 인공지능 알고리즘에 대한 법적 의무

항목	요구사항
훈련 데이터	위험 시나리오를 고려하고 충분히 광범위한 데이터셋에 기반해 훈련하는 등 안전을 합리적으로 보장할 것. 시스템의 사용이 금지된 차별을 수반하는 결과로 이어지지 않도록 합리적인 조치를 취할 것. 제품과 서비스를 사용하는 동안 사생활과 개인정보를 적절히 보호할 것
기록과 데이터의 보존	인공지능 시스템의 훈련 및 테스트에 사용된 데이터셋의 특성 및 선택 절차에 대하여 정확히 기록할 것, 정당한 경우 데이터셋 그 자체를 보존할 것, 시스템의 구축/테스트/검증에 사용된 프로그래밍 및 훈련의 방법론에

	대하여 문서화할 것
정보 제공	시스템의 역량과 한계, 특히 시스템이 의도한 목적, 의도한 대로 기능할 것으로 기대할 수 있는 조건 및 특정 목적을 달성하는 데 예상되는 정확도 수준에 대한 명확한 정보를 제공할 것, 시스템과 상호작용하는 시민들에게 사람이 아니라 인공지능 시스템이라는 사실을 알릴 것
견고성 및 정확성	시스템의 모든 생애주기 단계에서 견고하고 정확하거나 최소 자기 정확성의 수준을 올바르게 반영하도록 보장할 것, 결과를 재현할 수 있도록 보장할 것, 시스템의 모든 생애주기 단계에서 오류 또는 불일치를 적절히 처리할 수 있도록 보장할 것, 공개적인 공격 및 데이터 또는 알고리즘 자체를 조작하려는 교묘한 시도 모두에 대해 탄력성을 가지고 완화 조치를 취하도록 보장할 것
인적 감독	시스템의 결과물이 사전에 사람에 의해 검토되고 검증되지 않은 경우 효력이 없는 경우(사회보장급여 신청에 대한 거부 등), 시스템의 결과물에 즉시 효력이 발생하지만 사후 인적 개입이 보장되는 경우(신용카드신청에 대한 거부 등), 설계단계에서 시스템의 작동을 제약하는 경우(무인자동차의 센서 가시성이 저하되는 경우 작동을 중지시키거나 선행 차량과 일정 거리를 유지하는 등)
원격 생체인식 등 특정 애플리케이션에 대한 특별 요건	GDPR은 자연인을 고유하게 식별하려는 목적으로 생체인식 정보를 처리하는 것을 원칙적으로 금지하고 상당한 공익상의 이유로 법률에 따라 처리하는 경우 등에서만 예외적으로 인정함

- 고위험 인공지능에 대한 요구사항이 준수되는지 검증하고 보장하기 위해 객관적이고 사전적인 적합성 평가(Prior Conformity Assessment)를 의무적으로 실시하도록 함. 사전 적합성 평가는 인공지능 시스템에 대한 테스트, 검사 또는 인증 절차로 이루어지며 개발 단계에서 사용되는 알고리즘과 데이터셋에 대한 점검이 포함됨. 사전 적합성 검사 부적합 판정 시 인공지능 시스템을 재교육하도록 하고, 중소기업에 대해 온라인 검사 도구 등을 지원하도록 함. 시스템으로부터 부정적 영향을 받은 당사자에 대한 효과적인 사법적 보상을 보장하도록 함

5. 싱가포르 정부 <위험평가 매트릭스>[36]

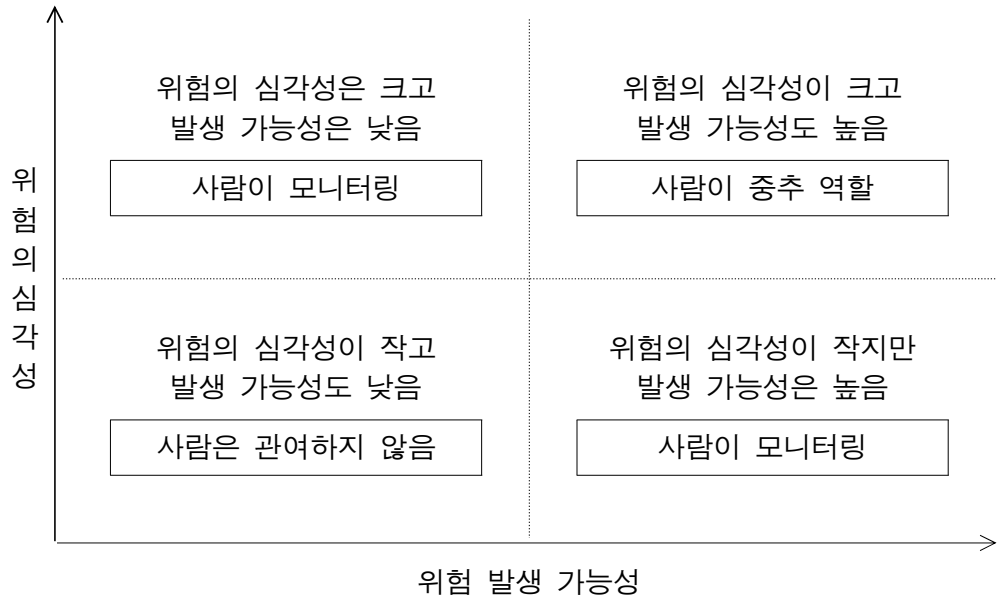
○ 싱가포르 정보통신규제청은 한 개인에 대한 기관의 의사결정 결과로 개인에게 해를 끼칠 확률과 심각도를 분류하는 매트릭스를 제안하여 평가에 활용

- 피해의 정의와 확률과 심각도의 계산은 상황과 부문에 따라 다름

○ 위험 매트릭스를 바탕으로 의사 결정에 필요한 인간 참여 수준 식별

- 예를 들어 안전이 가장 중요한 시스템의 경우라면 기관은 사람이 AI 시스템을 통제할 수 있도록 해야 하며 때에 따라서는 안전하게 종료할 수 있도록 해야 함

【싱가포르 정부】 위험평가 매트릭스



36) 한국정보화진흥원 (2019a). 앞의 문서 중 [참고4] 참조.

Ⅳ. 인공지능 모범 정책

1. 인공지능 영향평가

○ 인공지능이 기본권을 침해하는 영향을 완화하기 위하여 세계 여러 기구가 다양한 방식의 인공지능 영향평가를 도입하였거나 도입을 검토 중이며, 인권영향평가는 가장 효과적인 영향평가 방법으로서 권장되고 있음

○ 유엔 의사표현의 자유 특별보고관(David Kaye)은 2018년 보고서에서 인권에 기반한 인공지능 기술을 위하여 인권영향평가 또는 공공기관 알고리즘 영향평가의 실시를 각국 정부에 권고함[37)]

> 유엔 의사표현의 자유 특별보고관 권고 (2018)
>
> 62. 인공지능 시스템이나 응용 프로그램을 구하거나 사용할 때, 국가는 공공 부문 기관들이 지속적으로 인권의 원칙을 보장하도록 해야 한다. 그 중에서도 인공지능 시스템의 조달 및 사용 이전에 공공의 협의를 수행하고 인권영향평가 또는 공공기관 알고리즘 영향평가의 착수를 포함한다. 특히 인종 및 종교적 소수자, 정치적 반대 그룹이나 활동가에게 이런 기술이 미칠 수 있는 불평등한 영향에 더 신경을 써야 한다. 인공지능 시스템을 정부에서 사용하는 경우 외부의 독립적인 전문가로부터 정기적인 감사를 받아야만 한다.
>
> 63. 국가는 인공지능 시스템의 민간부문에서의 설계, 보급 및 실행에 있어서 인권이 중심에 올 수 있도록 해야만 한다. 이는 인공지능 영역에 대해 현 규제, 특히 개인정보보호 규제를 갱신하고 적용하는 것을 포함하며, 기업에 영향평가와 인공지능 기술에 대한 감사를 실시할 것을 요구하고 효과적인 외부 책임 메커니즘을 보장하도록 설계된 규제 혹은 공동 규제 체제의 추진을 포함한다. (하략)

37) Special Rapporteur on Promotion and protection of the right to freedom of opinion and expression (2018). “Report of the Special Rapporteur on Promotion and protection of the right to freedom of opinion and expression”, UN문서 A/73/348 (2018. 8. 29).

○ 유럽평의회는 인공지능 인권영향평가 회원국 권고를 추진중임(자동화된 개인정보 처리 및 인공지능의 인권 문제 전문위원회, 2019년 11월 초안 발표)

- 유럽평의회 인권위원장은 2019년 보고서에서 인권영향평가의 실시를 요구함[38]. 즉, 유럽평의회 회원국은 인권영향평가 수행을 위한 법체제를 수립할 것. 인권영향평가는 GDPR 개인정보 보호 영향평가 등 다른 영향평가와 유사한 방식으로 실시되어야 함. 인권영향평가는 인공지능 시스템을 검사하여 인권에 미치는 영향 및 위험성을 발견하고 조치하고 규명해야 함. 특히 공공기관은 인권영향평가의 공표나 수행이 가능하지 않는 공급자로부터 인공지능 시스템을 조달해서는 안 됨

○ 인공지능 영향평가에서 자주 참조되는 유럽연합 개인정보 보호법(GDPR)의 개인정보 보호 영향평가의 경우 다음의 고위험 개인정보 처리에 대하여 의무적인 영향평가를 실시하도록 함[39].

- '아동' 등 정보주체와 개인정보 처리자 간에 불균등한 권력관계에 처한 데이터를 처리하는 경우에는 의무적인 개인정보 보호 영향평가의 대상임

유럽연합 GDPR의 개인정보 보호 영향평가 의무화 대상

▸평가나 점수화. 특히 신용평가, 질병 예측을 위한 유전자 검사, 맞춤 마케팅 등 사람에 대한 프로파일링 및 예측의 경우

▸법적 혹은 이와 유사한 중대한 효과를 미치는 자동화된 결정

▸체계적인 감시. 특히 정보주체가 인지하지 못하는 사이에 공공장소 등에서 개인정보가 수집, 이용되는 경우

▸민감정보 또는 통신비밀, 위치정보, 금융정보 등 매우 사적인 데이터

▸정보주체의 수, 처리되는 데이터의 양과 범위, 데이터 처리 행위의 지속성 및 영구성, 처리행위의 지리적 범위 등에서 대규모로 처리되는 데이터

38) Council of Europe Commissioner for Human Rights (2019). 앞의 문서.

39) Article 29 Working Party (2016). "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".

▸데이터셋의 연계 또는 결합. 정보주체의 합리적 기대를 벗어나 다른 처리자에 의해, 다른 목적을 위해 처리되는 둘 이상의 데이터 처리의 경우

▸취약한 정보주체에 대한 데이터. 아동, 노동자 및 정신질환자, 망명신청자, 노인, 환자 등 처리자와 정보주체의 불균등한 권력관계에 처한 경우

▸신기술의 혁신적인 사용 또는 기술적, 조직적으로 새로운 솔루션의 적용 시에는 영향평가의 의무적 실시. 예를 들어 물리적 접근통제를 위해 지문이나 얼굴인식 기술을 사용하는 경우

▸처리 자체가 정보주체의 권리 행사 및 서비스접근이나 계약체결의 중단을 낳는 경우

○ 유럽연합 <인공지능 공공조달 백서>는 사전 위험 영향평가를 공공기관에 권장하였으며, 이때 실시할 수 있는 영향평가의 유형으로는 데이터윤리 영향평가, 법령준수 평가, 책임성 영향평가, 보안위협 평가, 사회환경 영향평가를 제시함

○ 영국 공직생활윤리위원회는 공공기관 인공지능에 대하여 의무적인 영향평가 실시 및 공개 제도 마련을 영국 정부에 권고함. 공공서비스 인공지능 공급자는 자체적인 영향 평가를 하고 시스템 설계에 있어 위험성 완화조치를 취해야 함[40)]

(정부/국가기관/규제기관에 대한 권고)

7. 의무적 영향평가 및 공개
- 정부는 인공지능이 공공표준에 미치는 잠재적 영향에 대한 평가를 현행 절차에 통합하는 방안을 검토해야 함. 이러한 평가는 의무적으로 실시하고 공개되어야 함

(공공 및 민간의 공공서비스 공급자에 대한 권고)

9. 공공표준에 대한 위험성 평가
- 공공서비스 공급자는 인공지능 시스템이 공공표준에 미칠 잠재적 영향을 평가하고, 시스템 설계가 공공표준에 미칠 위험을 완화하는지 확인해야 함. 인공지능 시스템 설계를 변경할 때마다 표준에 대한 검토가 이루어져야 함

40) The Committee on Standards in Public Life (2020). 앞의 문서.

- 영국 공직생활윤리위원회는 공공기관 인공지능에 의무적 영향평가를 적용해야 할 필요성에 대하여 4가지로 설명함

① 인공지능 시스템 관리 미흡은 공공기준 훼손 우려가 있음. 영향평가는 공공기관에 자기관 인공지능의 위험 수준을 인식시키고 그에 따른 위험관리를 실시하도록 함

② 인공지능에 대한 경험이 부족한 공공기관에게 영향평가가 데이터 편향 등 친숙하지 않은 위험성 문제를 다룰 수 있도록 함

③ 영향평가는 책임성 측면에서 중요함. 영향평가가 자기관 인공지능의 위험성을 인식하고 완화 조치를 취하도록 하여 공공기관의 적절한 책임성을 구현할 수 있음

④ 인공지능 기술은 국민 다수에 영향을 미치는 바 영향평가에서 그 이해관계 및 권리 보장 여부를 확실히 할 필요가 있음. 이때 영향평가는 상당한 주의 의무에 갈음될 수 있음

- 더불어 공공기관 인공지능 영향평가의 3가지 핵심 요소로서 다음을 제시함 △ 공공기관 인공지능 배치 전과 후에 영향평가를 의무적으로 실시할 것 △ 당사자 공공기관으로부터 분리된 제3자가 영향평가를 실시할 것 △ 영향평가의 결과는 공개되어야 함

ㅇ 영국 정부 인공지능 조달지침은 조달 절차 준비 및 계획 단계에서 인공지능 영향평가를 수행하도록 하고, 각 조달 절차에서 영향평가의 결과가 반영되는지 반복해서 평가하도록 함

- 인공지능 조달지침은 영향평가 실시에 있어 기존의 개인정보 보호 영향평가 및 평등 영향평가를 참고하여 실시할 것을 제안하고 특히 다음과 같은 영향을 유의하여 평가하도록 함

영국 정부 인공지능 조달지침 영향평가 항목

▸인공지능 시스템에 대한 사용자 요구사항과 그 공익

▸인공지능 시스템의 인적 및 사회 경제적 영향 - 이는 인공지능이 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

ㅇ 한편, 영국 앨런튜링 연구소는 공공부문에 이해당사자 영향평가(Stakeholder Impact Assessment)를 제안함[41].
이 평가는 공공기관들이 인공지능으로 영향을 받는 이해당사자를 확인하고 그 공정성 및 바람직한 결과물을 분석하며 인공지능 시스템이 개인과 사회에 미칠 수 있는 영향에 대해 검사한다는 장점이 있음

ㅇ 특히 캐나다 정부는 2019년 자동화된 의사결정 지침(훈령)을 발표하여 공공기관에서 자동화된 의사결정에 사용하는 인공지능의 도입 요건을 법규화하면서 알고리즘 영향평가의 사건 실시 및 공개를 의무화함. 이때 시스템 생산 전 각 공공기관은 알고리즘 영향평가(Algorithmic Impact Assessment, AIA))를 완료하고 각 시스템의 기능 또는 범위를 변경할 때마다 평가를 갱신하도록 하였음

ㅇ 호주 국가인권위원회는 2019년 <인권과 기술> 토론서에서 호주 정부에 인권영향평가의 개발 및 법규화를 제안함[42]

- 인공지능 정보 기반 의사결정과 관련하여 호주에 적용되는 모든 표준이 인권 준수에 대한 지침을 포함할 것과, 인권 중심 설계(human rights by design) 및 자율적·법적 인증제도를 또한 제안함

41) The Alan Turing Institute (2019). 앞의 문서
42) Australian Human Rights Commission (2019). “Human Rights and Technology : DISCUSSION PAPER” .

2. 투명한 정보공개와 참여

○ 공공부문 인공지능의 투명성을 보장하기 위한 각국의 노력이 계속되고 있음

- 유럽연합 집행위원회가 채택한 <신뢰가능 인공지능 가이드라인>은 인공지능의 투명성 보장 측면에서 원칙적으로 설명가능성을 요구하였음. 더불어 생애주기 전반에 걸쳐 인공지능 평가에 이해관계자가 참여할 것을 권장함. 이어 <인공지능 공공조달 백서>는 공공조달 인공지능 시스템이 추적가능하고 설명가능하고 이해관계자를 수용하도록 함

- 미국국립표준기술연구소 또한 2020년 8월 설명가능 인공지능 시스템을 위한 원칙을 제안하는 등[43] 설명가능 인공지능 발전을 위한 각국의 노력 또한 계속되고 있음

○ 최근 암스테르담과 헬싱키 시는 시민들에 알고리즘 등록부를 공개하기 시작함

【사례】 암스테르담과 헬싱키 시, 알고리즘 등록부 공개[44]

▸네덜란드 암스테르담과 핀란드 헬싱키 시는 인공지능의 투명성을 최대한 보장하고 시민의 신뢰를 확보하기 위해 시에서 사용하는 인공지능 알고리즘에 대해 등록하고 공개하는 '알고리즘 등록부'를 2020년 9월 공개함

▸알고리즘 등록부는 각 인공지능 시스템의 △훈련 데이터셋에 대한 정보 △데이터 처리에 대한 정보 △차별 방지에 대한 정보 △인간 감독에 대한 정보 △위험성에 대한 정보 등을 읽기 쉬운 평문으로 공개함

▸또한 알고리즘 등록부는 시에서 사용하는 알고리즘의 도입을 책임지는 공직자의 이름, 부서 및 연락처를 공개하고 시민들이 의견을 제출할 수 있도록 함

43) National Institute of Standards and Technology (2020). "Four Principles of Explainable Artificial Intelligence". <https://www.nist.gov/system/files/documents/2020/08/17/NIST%20Explainable%20AI%20Draft%20NISTIR8312%20%281%29.pdf>

44) 암스테르담 알고리즘 등록부 <https://algoritmeregister.amsterdam.nl/en/ai-register/>;

○ 특히 공공기관의 경우 대국민 정보공개 의무가 있다는 점에서 투명성 의무에 대한 법적인 판단이 내려지고 있음

【사례】 공공기관 인공지능 의사결정과 투명성[45)]

▸네덜란드 사회복지 위험발견시스템(SyRI)은 중앙정부 및 지자체가 본래 분리보관되어 있던 데이터들을 광범위하게 결합하여 이를 비공개 인공지능 "위험 모델"에 기반해 분석 후 부정수급 소지가 있는 사람들을 발견하려는 시스템이었음

▸네덜란드 헤이그 지방법원은 2020년 2월 SyRI 관련 법률의 프라이버시 침해 보호조치가 충분치 않고 그 작동 원리에 대한 "투명성이 중대하게 결여되어 있다"며 사용 중단을 명령함. 법원은 이 시스템이 추구하는 사회복지 부정수급자 발견이라는 목표가 사생활권 침해와 비례적이지 않아 위법하다고 판시함

○ 영국 정부는 <공공부문 인공지능 활용 가이드>에서 설명가능성과 투명성을 함께 보장하도록 함

- 영국 개인정보 보호 감독기구(ICO)는 앨런튜링 연구소와 함께 <인공지능 의사결정에 대하여 설명하기> 지침 초안을 발표함[46)]

- 영국 공직생활윤리위원회는 정부가 공공기관의 인공지능 시스템 관련 신고 및 정보공개에 관한 명확한 지침을 마련할 것을 제안함

- 이후 영국 정부 데이터 윤리 및 혁신 센터는 알고리즘 의사결정 편향성을 검토하면서 이를 완화하기 위한 조치로 투명성 의무가 되는 정부 알고리즘 의사결정 시스템의 범위를 탐색함[47)]

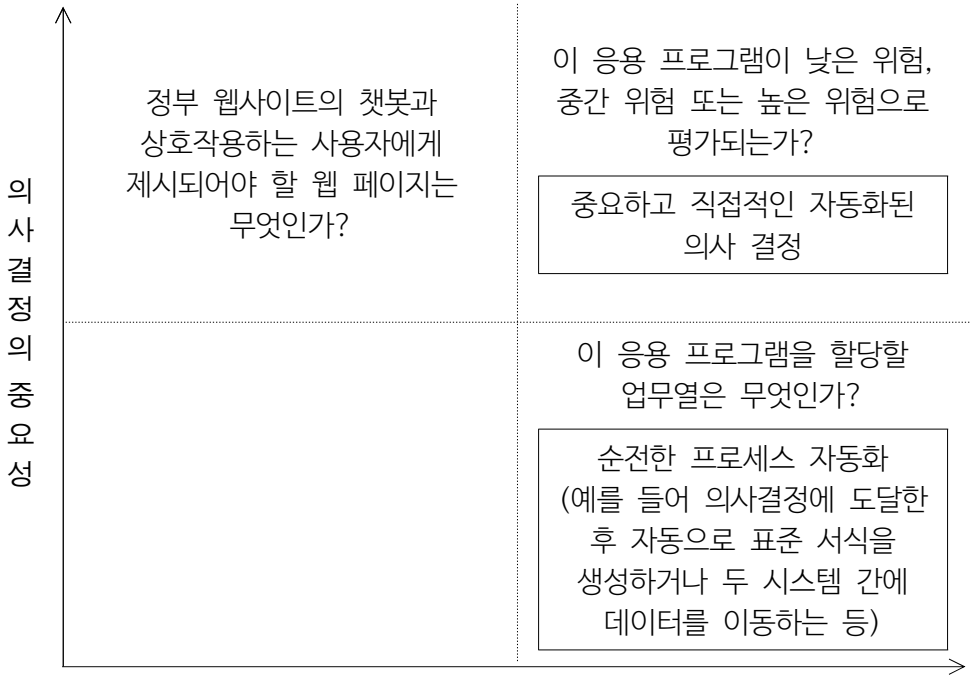
헬싱키 알고리즘 등록부 <https://ai.hel.fi/en/ai-register/>; 관련 언론보도 <https://venturebeat.com/2020/09/28/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/>

45) 가디언 관련 보도 <https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules>; 유엔 빈곤과 인권에 관한 특별보고관 보도자료 <https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25152&LangID=E>; 공익소송단 <https://pilpnjcm.nl/en/landslide-victory-in-syri-case-dutch-court-bans-risk-profiling/>

46) ICO and the Alan Turing Institute (2019), "Explaining decisions made with AI: Draft guidance for consultation".

47) Centre for Data Ethics and Innovation (2020) "Independent report: Review into bias in algorithmic decision-making".

【영국 정부】 투명성 의무 대상 정부 알고리즘 의사결정 시스템의 범위(예시안)



○ 캐나다 자동화된 의사결정에 대한 지침(훈령)에서는 투명성 보장을 위한 규정들 두고 △자동화된 의사결정 이전에 공지하고 △자동화된 의사결정 후에는 설명하며 △그 구성 요소에 대한 접근 권한을 보장하고 △가능한 소스 코드를 공개할 것 등을 명시함

○ 호주 국가인권위원회는 정부가 인공지능 정보 기반 의사결정 시스템의 도입을 계획할 경우 가장 영향을 받을 가능성이 높은 사람들에 초점을 맞춘 공청회를 개최할 것을 제안함[48)]

48) Australian Human Rights Commission (2019). 앞의 문서.

3. 정보주체 권리 보호

ㅇ 인공지능의 자동화된 의사결정에 있어서 그 예측곤란성과 자율성을 명분으로 국민 권리구제의 공백이 발생할 우려가 있음. 이에 정보주체의 권리 보호 및 구제를 위한 다양한 정책 마련이 요구됨

- 공공기관이 운용하는 인공지능의 경우 그 의사결정으로 영향을 받는 사람들의 권리를 명확하게 정의하고 헌법이 보호하는 적법절차를 법적으로 보장할 필요가 있음

- 특히 최근 인공지능 규범은 일반 시민에 대한 정보공개 원칙과 구분하여 인공지능 의사결정으로 영향을 받는 사람들(affected individuals) 및 정보주체의 권리를 정의하고 보장하려는 추세로 발전해 옴

ㅇ 유럽연합은 인공지능의 특성으로 인해 그 의사결정의 관련 법률 준수 여부를 검증하거나 효과적인 이행에 지장을 초래할 수 있고, 당국과 피해자들은 의사결정 과정을 추적·검증하거나 사법적 수단을 비롯한 권리구제 접근에 어려움을 겪을 수 있다고 우려함

- 이에 '신뢰가능 인공지능 가이드라인'에서 신뢰가능 인공지능 윤리 실현을 위한 인공지능 생애주기 평가에 이해관계자 참여를 제안함. 나아가 후속 <인공지능 백서>에서는 영향을 받는 이해당사자에 대한 정보제공 의무를 명시함

- 유럽연합은 특히 개인정보 보호법(GDPR)에서 정보주체가 프로파일링 등, 본인에 관한 법적 효력을 초래하거나 이와 유사하게 본인에게 중대한 영향을 미치는 자동화된 처리에만 의존하는 결정의 적용을 받지 않을 권리를 보장하고 있음(제22조). 정보주체는 프로파일링 등 중대하고 유일한 자동화된 의사결정의 유무에 대하여 사전에 통지를 받을 수 있어야 하며, 이때 관련 논리(logic)에 관한 유의미한 정보와 그 같은 처리가 본인에게 미치는 중대성 및 예상되는 결과에 대한 설명이 이루어져야 함(제13조). 또한 중대하고 유일한 자동화된 의사결정의 적용을 받는 경우라 하더라도 인적개입을 획득할 수 있는 권리, 의사를 표현할 권리, 이러한 평가 이후 도달한 결정에 대한

설명을 획득할 권리, 해당 결정에 이의를 제기할 권리 등, 적절한 안전조치를 보장받음. 이러한 의사결정에서 ‘아동’은 제외하도록 함(전문71)

ㅇ 유럽평의회 인권위원장은 회원국 인공지능 시스템의 발전과 구현에 있어 특히 중대하게 영향을 받는 이해당사자들(affected individuals)의 의견수렴, 정보제공, 권리구제를 요구함[49)]

- 인권위원장은 회원국들이 인공지능으로부터 그 권리에 부당한 영향을 받을 위험성이 큰 ‘아동’ 등의 사회집단에 대한 차별을 방지해야 한다고 강조함

【유럽평의회 인권위원장】 인권 규제 준수를 위한 주요 실행 영역

1 인권영향평가 - 다른 영향평가와 유사한 방식으로 인권영향평가를 실시하는 법체제를 수립하고 공공기관은 이를 조달에 반영해야 함
2 공개적인 의견수렴 - 인권영향평가나 조달 등 단계별로 인공지능 시스템의 운영, 기능, 영향 등 세부사항을 공표하고 의견을 수렴해야 함
3 민간기업의 인권 기준 준수 - 모든 인공지능 운용자는 인권 원칙 준수 여부를 확인하고 공표해야 함. 투명한 인권 실사 절차 수용으로 인공지능 시스템의 인권 위험을 확인할 수 있어야 함
4 정보제공과 투명성 - 인공지능 의사결정의 대상 개인들은 이에 대해 고지받고 지체없이 전문가의 조력을 선택할 수 있어야 함. 인공지능 시스템에 인권적 검토나 정밀검사가 허용되어야 함
5 독립적인 감독 - 인공지능의 인권 준수에 대한 독립적이고 효과적인 감독을 위하여 법제도 마련. 독립적인 기구가 준수 여부를 조사하고 영향을 받은 개인 진정을 처리하고 인공지능 시스템의 성능 발전에 따른 정기적인 검토를 수행할 수 있도록 해야 함
6 차별금지 및 평등 - 인공지능 시스템은 차별을 방지하기 위하여 높은 수준의 정밀검사를 받아야 하며, 특히 인공지능으로부터 그 권리에 부당한 영향을 받을

49) Council of Europe Commissioner for Human Rights (2019). 앞의 문서.

위험성이 큰 사회집단(아동, 노인, 장애인 등)에 대한 차별을 방지해야 함. 이 원칙은 특히 법집행기관의 프로파일링을 방지하기 위해 중요함

7 개인정보 보호 및 프라이버시
- 인공지능 시스템은 개인정보 처리에 대한 법적 근거와 공정하게 비례적이어야 함. 인공지능 시스템이 민감정보를 처리할 경우 높은 수준의 안전조치를 적용해야 함

8 표현의 자유, 집회결사의 자유, 노동권
- 기술적인 독점 형성을 방지하여 인공지능 전문성과 권한이 집중되고 정보의 자유로운 유통에 부정적인 영향이 미치지 않도록 해야 함. 회원국은 인공지능 발전으로 인한 일자리 창출과 실업의 수치와 유형을 추적해서 실업을 완화해야 함

9 권리구제
- 인공지능 시스템은 언제나 인적 통제 하에 속해야 함. 인공지능 인권침해에 대한 책임성과 책무성은 언제나 자연인과 법인이 감당해야 함. 최소 인적 개입을 보장받을 수 있어야 함. 효과적인 권리구제가 시행되어 인공지능 시스템의 결과로 인한 피해를 시정할 수 있어야 함

10 '인공지능 리터러시' 증진
- 인공지능 관련 문제를 자문할 수 있는 정부내 협의기구의 설립을 검토해야 함

○ 유엔 의사표현의 자유 특별보고관은 인공지능 시스템의 도입에 있어 권리주체 보호를 위해 취해져야 할 절차들을 제안함[50)]

▸인공지능 시스템의 활용 전후 및 활용 과정 중에 인권 영향 평가를 할 것(53. Human rights impact assessments)

▸인권 단체들과 함께 외부에서 감사하고 협의할 것(55. Audits)

▸개인이 선택할 수 있도록 고지와 동의 절차를 갖출 것(58. Individual autonomy; 59. Notice and consent)

▸인권침해를 종식시키기 위한 효과적인 구제 절차를 마련할 것(60. Remedy)

- 특히 특별보고관은 인공지능 시스템으로부터 영향을 받은 개인들에 대한 구제 수단이 확보되어야 한다고 강조함

50) Special Rapporteur on Promotion and protection of the right to freedom of opinion and expression (2018). 앞의 문서.

70. 개인 사용자들은 인공지능 시스템의 반인권적 영향의 구제 수단에 접근 할 수 있어야만 한다. 기업들은 인공지능에 따라 처리되는 시스템에 부과되어 나타나는 모든 사용자들의 불만 및 항의에 적시에 대응하기 위해 사람에 의한 평가 및 구제 시스템을 두어야 한다. 인공지능 시스템에 불만이 제기되고 구제가 요청된 횟수에 대한 데이터 뿐만 아니라, 이용 가능한 구제책의 종류와 효과성에 대해서도 주기적으로 공개 되어야 한다.

ㅇ 유엔 인종차별철폐위원회는 2020년 12월 17일 <법집행관리의 인종적 프로파일링 방지와 대응을 위한 일반 권고>에서 (A) 법제도・정책 관련 조치 (B) 인권 교육・훈련 관련 조치 (C) 채용 조치 (D) 지역사회 정책 (E) 데이터 조치 (F) 책임성 조치 (G) 인공지능 정책으로 나누어 각국 정부에 권고함[51)]

- 특히 사회 각 영역에서 인공지능의 인종적이고 차별적인 의사결정을 방지하기 위한 법제도・정책 관련 조치로서
△법집행관리의 인종적 프로파일링을 정의하고 금지하는 법과 정책을 수립 및 실행할 것. 인종적 프로파일링을 유발하거나 촉진할 가능성이 있는 현행 법제도에 대한 개선에 착수할 것 △인종프로파일링을 방지하기 위한 법집행기관 검문검색 지침을 개발하고, 내외부 모두에서 효과적이고 독립적인 감독 체제와 위반에 대한 제재 조치를 포함할 것 △모든 국가는 인종차별로부터 모든 사람을 보호하고 구제할 것을 보장하고 이러한 피해로부터 적정한 배보상을 구할 수 있어야 함 △국가는 피해자 중심적인 접근으로 관련 부처, 지자체, 시민단체, 상호교차적 차별을 경험하는 집단을 대표하는 사람들, 국가인권기구들과 협력증진을 통하여 지원 서비스를 효과적으로 조율할 것을 권고함

ㅇ 영국 정부는 여러 인공지능 규범에서 그 대상이 되는 시민의 권리를 정의하고 보장하도록 함

51) Committee on the Elimination of Racial Discrimination. (2020) “General recommendation No. 36 (2020) on preventing and combating racial profiling by law enforcement officials“. UN문서 CERD/C/GC/36 (2020. 12. 17).

- <공공부문 인공지능 활용 가이드>는 인공지능 활용 의사결정이 개인에게 법률적으로 영향을 줄 수 있는 경우라면 반드시 유럽연합 GDPR 및 영국 개인정보보호법 규정에 따라 정보주체 보호 조치를 필수적으로 갖추도록 함[52)]

▸자동화된 의사결정 절차에 대하여 구체적이고 쉽게 접근할 수 있는 정보 제공
▸인공지능 의사결정에 대해 인간이 검사하고 결정을 변경하는 등 개입할 수 있는 명확한 방안 마련

- 공공부문을 위한 <인공지능의 윤리와 안전을 고려한 시스템 설계·구현 가이드> 또한 인공지능 모델이 처리된 방법과 근거 등을 영향을 받는 이해당사자들(affected stakeholders)에게 공개하는 투명성 원칙을 명시함[53)]

- 공직생활윤리위원회는 공공서비스 인공지능 공급자로 하여금 공공서비스 대상자인 시민에게 권리를 안내하고 이의제기 방법을 알리도록 함[54)]

(공공 및 민간의 공공서비스 공급자에 대한 권고)
14. 이의제기 및 배상방법 안내
- 공공서비스 공급자는 시민에게 그들의 권리와 인공지능 기반 결정에 대해 이의제기하는 방법을 알려야 함

○ 캐나다 자동화된 의사결정 지침(훈령)에서는 공공기관 인공지능의 의사결정 후에 영향을 받는 개인들에게 그 결정이 내려진 방법과 이유에 대하여 이해가능하게 설명하도록 함

○ 호주 국가인권위원회는 정부에 인공지능 정보 기반으로 의사결정에 이를 경우 그 영향을 받은 사람에게 관련 정보를 제공하고, 의사결정의 설명가능성을 보장하는 법안을 정부에 마련하도록 제안함

52) Government Digital Service and Office for Artificial Intelligence (2019a). 앞의 문서.
53) Government Digital Service and Office for Artificial Intelligence (2019b). 앞의 문서.
54) The Committee on Standards in Public Life (2020). 앞의 문서.

- 인공지능 정보 기반 의사결정에 영향을 받은 사람에게 제공될 설명은 의사결정 사유를 포함해서 개인 또는 관련 기술 전문가가 의사결정의 기반을 이해하고 이의 제기가 가능한 근거를 이해할 수 있어야 함. 개인의 인권을 침해할 수 있는 의사결정에 대해 합리적인 설명을 제공하지 않는 경우 인공지능 정보 기반 의사결정 시스템을 도입해서는 안 됨[55)]

ㅇ 종합해 보면, 공교육 인공지능의 경우 알고리즘을 이용한 의사결정이 이루어지기 전에 국민 일반에게 정보와 설명을 공개하고 특히 공청회 등의 방식으로 영향을 받는 주체들의 의견을 수렴하는 절차가 보장되어야 함. 더불어 공공과 민간을 아울러 정보주체에게 법적이거나 상당한 영향을 미치는 자동화된 의사결정의 경우 영향을 받는 주체에게 그 사실과 주요 로직, 장래의 영향에 대하여 사전에 통지하고 거부권을 보장할 필요가 있음. 이러한 처리에서 아동은 제외하는 것이 바람직함

- 한편 인공지능으로 자동화된 의사결정으로 공공 처분이 이루어지는 경우, 청문권 문서열람권 결정의 이유제시요구권 등 헌법상 적법 절차를 보장해야 함. 더불어 공공과 민간을 아울러 정보주체에게 법적이거나 상당한 영향을 미치는 자동화된 의사결정의 경우 정보주체의 동의나 법률 계약에 의하여 적법하게 처리하는 경우라 하더라도 인적개입 요구권, 의견 진술권, 이의제기권 및 권리구제를 보장해야 함

55) Australian Human Rights Commission (2019). 앞의 문서.

V. 제언

1. 연구의 개요

○ 본 연구에서는, 인공지능 관련 국내외 정부기관들의 거버넌스 정책과 기준을 사례로 분석하여 공교육에 도입되는 인공지능 알고리즘의 위험성을 평가하고 위험의 수준별 기준을 구축하고 관리하는 방법을 포함하는 공교육 적용 인공지능 알고리즘의 공공성 확보 방안을 제시하였음

- Ⅱ장에서는 인공지능 윤리, 공공기관 인공지능 윤리, 관련 법령 등 인공지능 알고리즘 관련 일반 규범을 검토함
- Ⅲ장에서는 캐나다 정부 〈자동화된 의사결정에 대한 지침〉, 뉴질랜드 정부 〈위험성 매트릭스〉, 독일 정부 데이터윤리위원회 〈알고리즘 시스템 위험도 피라미드〉, 유럽연합 〈위험기반 접근법〉과 싱가포르 정부 〈위험평가 매트릭스〉 등 해외 여러 국가에서 인공지능 알고리즘 시스템의 위험성을 평가하고 위험 등급별 관리를 도입하였거나 추진 중인 현황을 검토함
- Ⅳ장에서는 그밖의 인공지능의 모범 정책으로 인공지능 영향평가, 투명한 정보공개와 참여 보장, 정보주체의 권리 보장 정책 등에 대하여 검토하고, 국내 적용 방안을 제시함
- 마지막으로 Ⅴ장에서는 이상의 연구 검토 결과를 토대로 공교육에 적용되는 인공지능 알고리즘의 공공성을 확보하기 위한 종합적 제언을 도출함

2. 공교육 적용 인공지능 알고리즘 거버넌스를 위한 제안

가. 학습자의 성장을 최우선 원칙으로 하는 공교육 적용 인공지능 알고리즘 원칙 수립

- ㅇ 공교육 학교시스템에 포함되는 학생은 대부분 신체적으로나 정신적으로 성인에 이르기까지 발달하는 과정에 있는 인간이라는 특수성을 고려하여 공교육에 적용되는 인공지능 알고리즘 원칙을 고안해야 함

- ㅇ 미성년 학생과 관련된 공교육 적용 인공지능 알고리즘 원칙은 학생 성장의 원칙을 최우선 원칙으로 고안되어야 함을 제안함

나. 인공지능 알고리즘에 대한 개인적/사회적/기술적 차원의 통제력 확보 체제 구축

- ㅇ 개인적 차원 그리고 사회적 차원에서 공존의 대상으로 인공지능을 다루기 위해서 개인적 차원, 사회적 차원, 기술적 차원에서의 통제력 확보가 필요함

- ㅇ 개인적 차원의 통제력 확보 : 인공지능 알고리즘에 대한 개인적 차원의 통제력 확보를 위해서는 개인정보, 데이터, 알고리즘 등에 대한 기본 소양 교육이 필요함. 특히 학생이 정보주체로 현명하게 성장할 수 있도록 기여하는 교육프로그램이 제공되어야 함

- ㅇ 기술적 차원의 통제력 확보 : 인공지능 기술, 자동화 알고리즘, 개인정보/데이터의 기술적 거버넌스 확보를 위한 기술환경의 (재)설계가 핑요하며, 데이터의 설계 -> 데이터의 확보 -> 데이터의 검증 -> 데이터의 전처리 -> 알고리즘 모형의 선택과 적용 -> 기계학습의 과정과 결과 -> 처리 결과의 시각화와 표시 -> 알고리즘 모형의 업데이트 등 인공지능 알고리즘 전체 프로세스에 대한 통제력을 확보하기 위한 기술환경의 (재)설계가 필요함

- ㅇ 사회적 차원의 통제력 확보 : 공교육 정책을 기획, 실행, 평가하는 국가기관에서는 적절한 원칙, 정책, 규칙을 마련하고 발생하는 현상을 안정적으로 모니터링하는 프로세스와 추진 체제를 구축해야 함.

다. 인공지능 알고리즘의 공교육 적용을 위한 거버넌스 방안 제안

ㅇ 인공지능이 교육에 미칠 영향 평가 (인공지능 영향평가) 체제 구축

- 개인정보를 포함한 데이터의 자동화 처리 알고리즘과 심층기계학습에 기초한 인공지능 기술이 공교육 또는 교육 전체에 미칠 영향을 평가하기 위한 체제와 추진 주체가 마련되어야 함
- 인공지능 영향 평가의 원칙과 프로세스를 구축하고, 해당 영역의 인공지능 알고리즘 위험성 등급에 따른 처리 절차를 정의하고, 그 영향을 사전/사후 평가를 실행할 수 있는 체제가 마련되어야 함
- 인공지능영향평가는 인공지능 시스템을 검사하여 학생과 교원의 인권, 교수학습 과정과 결과에 미치는 영향 및 위험성을 사전적으로 발견하고 이에 대해 조치하고 규명해야 함. 특히 공교육 기관은 인공지능 영향평가의 수행 및 공표가 가능하지 않는 공급자의 경우 인공지능 시스템을 조달할 수 없도록 제도적인 안전장치가 마련되어야 함

ㅇ 인공지능 알고리즘 투명성 확보(정보공개와 참여) 체제 구축

- 공교육에 적용되는 인공지능 알고리즘의 신뢰가능성, 투명성, 설명가능성을 확보하기 위해 인공지능 시스템의 생애주기 전체에 걸쳐 추적가능하며 이해관계자의 참여가 가능한 평가 체제를 구축해야 함
- 이해관계자들이 해당 인공지능 모델의 데이터 처리 방법과 과정에 접근가능한 인공지능을 공교육 추진 기관이 도입할 수 있도록 조달 프로세스를 개선해야 함
- 공교육에 적용되는 인공지능 알고리즘의 자동화된 의사결정에 대한 가이드라인을 마련하여, 기계학습 데이터셋에 대한 정보, 데이터 처리 과정에 대한 정보, 차별 방지 방법, 인간 감독 주체에 대한 정보, 위험성(개인차원, 사회차원)에 대한 정보를 공개하도록 하는 체제를 마련해야 함

ㅇ 정보주체로의 성장을 지원할 역량 강화 프로그램 구축

- 공교육에 인공지능 알고리즘이 도입될 경우, 중대하게 영향을 받는 이해당사자들(affected individuals)이라고 할 수 있는 학생과 교원이 정보주체로 성장할 수 있는 역량 강화 프로그램 구축

- 공교육에 적용되는 인공지능의 경우, 알고리즘을 이용한 의사결정이 이루어지기 전에 학생(학부모 등의 법률 대리인)과 교원에게 데이터 처리 절차와 알고리즘의 모형 등에 대한 정보와 설명을 공개하고 특히 공청회 등의 방식으로 영향을 받는 주체들의 의견을 수렴하는 절차가 보장되어야 함.

- 정보주체로서의 역량을 키워갈 수 있도록 정보의 공개, 토론, 협의에 학생 참여를 보장해야 함

- 더불어 학생 뿐만 아니라 교원에게도 중요한 영향을 미칠 가능성이 있는 자동화된 의사결정의 경우, 영향을 받는 주체에게 그 사실과 주요 처리방법 등, 장래의 영향에 대하여 사전에 통지하고 거부권(미성년 학생일 경우에는 법적 대리인)을 보장할 필요가 있음

ㅇ 공교육에 적용되는 인공지능 알고리즘의 공공성 확보를 위한 거버넌스 체제 구축

- 공교육에 적용되는 인공지능 알고리즘 개발 전문기관 (회사, 연구소 등), 관련 정책을 실행할 교원 (전문교원, 교원단체 등), 학생과 학부모 등이 참여하는 협의체와 인공지능 알고리즘이 도입되는 개별 정책의 실행과정과 결과를 관리 감독할 수 있는 거버넌스 체제 구축

3. 서울형 공교육 적용 인공지능 알고리즘 공공성 확보 가이드라인 (안)

가. 공교육 적용 인공지능 알고리즘 원칙(안)[56]

▸성장 원칙[57]

학생은 신체적 정신적 성장의 과정에 있으며, 모든 의사결정은 이러한 성장의 지원에 기초해야 하며, 원칙들이 충돌할 경우에도 성장 원칙은 최우선 원칙으로 고려되어야 한다.

▸개인화 원칙

학생은 본인의 발달 정도에 따라 적응적인 프로그램을 제공받을 권리가 있으며, 이 과정에서 개인정보보호 원칙과 충돌할 경우에도 미성년 학생의 성장을 위한 원칙 즉, 학생이 성장해야 하는 교육적인 목적에 따른 개인정보 활용이 적극적으로 고려되어야 한다.

▸정보주체 원칙

학생은 정보주체로 성장할 수 있도록 지원받아야 하며, 미성년 단계의 학생이라고 하더라도 본인과 관련된 정보의 주체가 되어야 하고 정보주체로서의 의사결정은 법정 대리인 원칙과 조정되어야 한다.

▸개인정보보호 원칙

미성년 학생이라고 하더라도 개인정보가 보호되어야 하며, 개인정보에 기초하여 가공되거나 생성된 2차 정보 역시 개인정보의 범위 내에서 관리되어야 한다.

▸복지 원칙

미성년 학생은 복지의 관점에서 사회적으로 보호받고 돌봄의 대상이 되어야 한다.

▸법정 대리인 원칙

미성년 학생은 법정 대리인의 보호를 받을 수 있어야 한다. 단, 법정대리인과 미성년 학생의 이해와 의견이 충돌할 경우, 이러한 충돌을 조정하는 과정에서 미성년 학생을 자기정보의 주체로 고려되어야 한다.

---

56) 원칙은 EU ARTICLE 29 DATA PROTECTION WORKING PARTY(2009) Opinion 2/2009 on the protection of children's personal data (General Guidelines and the special case of schools), Adapted 11February 2009, 398/09/EN, WP 160을 기초로 구성함

57) EU의 2009년 개인정보보호 가이드라인의 학교 사례 the protection of children's personal data (General Guidelines and the special case of schools)에서 '성장 원칙'은 Best interest of the child, '복지 원칙'은 Protection and care necessary for the wellbeing

나. 공교육 적용 인공지능 알고리즘 위험성 평가 매트릭스(안)[58]

ㅇ 위험성 평가 과정에서 제기되어야 하는 질문

(1) 인공지능 알고리즘의 자동화된 의사결정이 미치는 영향의 범위는 어떠한가? 독립적이어서 영향이 확대되지 않는가? 확대되더라도 제한적인가? 그 범위가 광범위하여 제한할 수 없는가?

(2) 인공지능 알고리즘의 자동화된 의사결정이 발생시킬 부정적인 효과가 어떠한가? 발생할 가능성이 높은가? 가능성이 높지는 않지만 발생할 가능성은 있는가? 발생할 가능성이 거의 없는가?

(3) 인공지능 알고리즘의 자동화된 의사결정이 발생시킬 효과는 얼마나 심각한가? 매우 심각한가? 경우에 따라 심각할 가능성이 있는가? 심각할 가능성이 거의 없는가?

ㅇ 제기되어야 하는 질문에 대한 응답을 매트릭스로 구조화하면 아래 그림과 같음

발생가능성

발생가능성 높음 일반적인 작동 중에 발생할 가능성 높음			
발생가능성 있음 일반적인 작동 중에 발생할 가능성 있음			
발생가능성 거의 없음 일반적인 작동 중에 발생할 가능성이 낮지만 발생할 수는 있음			
영향 정도	**낮음** 영향이 미치는 범위가 독립적이며, 심각할 가능성 낮음	**중간** 영향이 미치는 범위가 제한적이며, 경우에 따라 심각할 가능성이 있음	**높음** 영향이 미치는 범위가 광범위하며, 매우 심각할 가능성이 높음

ofchildren, ‘프라이버시 원칙’은 Right to privacy , ‘법정 대리인 원칙’은 Representation으로 표기했다.

58) 뉴질랜드 정부의 위험성 매트릭스를 활용하여 수정한 버전임

다. 서울형 공교육 적용 인공지능 알고리즘 위험성 평가 프로세스(안) : 데이터의 수집/처리, 알고리즘 모형의 선택/기계학습

ㅇ 공교육에 적용되는 인공지능 알고리즘의 위험성 평가는 데이터의 수집, 처리, 결과의 출력에 이르는 인공지능 시스템의 전체 생에주기에 걸쳐 평가 체제를 구축해야 함

ㅇ 데이터의 측정/수집, 전처리 단계

(1) 사용할 데이터의 구조를 설계하는 초기 단계에서, 데이터를 정의하는 핵심 키워드/카테고리/토픽 등이 대상 데이터와 관련된 사용자(학생, 학부모, 교원, 행정직원 등)에게 미칠 영향을 위험성 평가 매트릭스에 준하여 평가해야 함.

(2) 데이터의 측정/수집이 필요한 데이터의 경우에는, 데이터의 측정/수집 과정에서 개인정보 보호, 본인 및 법적 보호자의 허가 및 허가의 방법 등이 적접하게 설계되었는지 그리고 적법하게 처리되었는지를 확인하는 절차를 평가해야 함

(3) 기존에 측정/수집된 데이터를 사용할 경우에는, 해당 데이터의 출처와 측정/소유/관리하는 과정에서 필요한 절차가 적법하게 진행되었는지를 확인하는 절차를 평가해야 함

ㅇ 알고리즘 모형의 선택 및 기계학습 적용 단계

(1) 알고리즘 모형을 선택할 경우, 내부의 구조와 데이터 처리 과정이 투명성/신뢰성/설명가능성의 원칙에 부합하는 알고리즘 모형이 선택되었는지를 평가해야 함

(2) 선택한 알고리즘을 선택한 데이터를 이용하여 기계학습을 시킬 경우, 기계학습 과정에서의 데이터 처리 과정이 투명성/신뢰성/설명가능성의 원칙에 부합하는 기계학습 방법이 선택되었는지를 평가해야 함

ㅇ 알고리즘 모형의 수정 및 업그레이드 단계

(1) 선택한 알고리즘 모형이 선택한 데이터를 이용한 기계학습의 결과로 수정 및 업그레이드될 경우, 수정 및 업그레이드된 알고리즘 모

형 결과물 역시 투명성/신뢰성/설명가능성의 원칙에 부합하도록 수정 및 업그레이드 되는지를 평가해야 함

ㅇ 데이터 처리 결과의 출력 및 피드백 단계

(1) 데이터 처리 결과의 출력 또는 처리 결과의 적용의 대상자가 미성년 학생일 경우, 미성년 학생에게 직접 적용하기 전 교원과 학부모에게 사전 공지 되어야 하며, 영향성평가위원회(가칭) 등의 결정에 따라 미성년 학생에게 직접 데이터 처리 결과를 공지하거나, 미성년 학생에게 직접 제공되는 서비스에 적용하지 않고 대상 학생의 교원 또는 학부모를 통해서 전달받도록 하는 대상 사용자에 따른 데이터 처리 결과의 출력 및 출력 방법을 평가해야 함

(2) 데이터 처리 결과의 출력 또는 처리 결과의 적용의 대상자가 미성년 학생일 경우와 성년 사용자일 경우 모두, 데이터 처리 결과 또는 처리 결과 적용 서비스에 적용된 데이터 처리 과정에 대한 설명을 요구할 수 있는 피드백 절차가 마련되어 있는지 평가해야 함

라. 서울형 공교육 적용 인공지능 알고리즘 위험성 평가 프로세스(안) : 알고리즘의 선택 및 적용

ㅇ 공교육에 적용되는 인공지능 알고리즘의 위험성 평가는 알고리즘의 선택 및 적용의 전체 생에주기에 걸쳐 평가 체제를 구축해야 함

ㅇ 공교육에 적용할 인공지능 알고리즘의 선택

(1) 공교육에 적용할 인공지능 알고리즘 모형은 내부의 구조와 데이터 처리 과정이 투명성/신뢰성/설명가능성의 원칙에 부합하는 알고리즘 모형을 사용해야 하며, 이와 같은 원칙에 부합하는 알고리즘 모형이 선택되었는지를 알고리즘평가위원회(가칭)을 구성하여 평가해야 함

ㅇ 공교육에 적용할 인공지능 알고리즘의 시범 적용 및 결과 평가

(1) 선택한 알고리즘을 공교육 현장에서 측정/수집되는 데이터를 이용하여 기계학습을 시킬 경우, 기계학습 과정에서의 데이터 처리 과정이 투명성/신뢰성/설명가능성의 원칙에 부합하는 기계학습 방법이 선택

되었는지를 평가해야 하며, 이를 위해 시범적용을 위한 테스트베드 환경이 마련되어야 함

(2) 위험성 평가 매트릭스를 이용한 영향 평가 결과 테스트베드에서의 시범적용이 필요하다고 결정될 경우, 정해진 기간동안 시범적용을 진행할 수 있는 테스트베드 환경이 제공되어야 함

(3) 인공지능 알고리즘의 공교육 현장 시범 적용을 위한 테스트베드 환경을 설치할 경우, 학교 현장과 동일한 학교 환경으로 구축되어야 하며 해당 공간으로 진입하기 전, 데이터의 측정/수집/처리/기계학습에의 적용/알고리즘 기능개선에 활용 등의 과정에 개인정보가 활용될 수 있음을 모든 사용자에게 공지해야 하며, 테스트베드 환경에서의 데이터 측정/수집/처리/기계학습에의 적용/알고리즘모형 개선 과정이 투명성/신뢰성/설명가능성의 원칙에 부합하는를 평가해야 함

ㅇ 인공지능 알고리즘의 공교육 현장 적용 및 피드백 절차

(1) 인공지능 알고리즘이 직접 공교육 현장의 사용자에게 제공될 경우, 인공지능 알고리즘이 적용되고 있다는 사실을 사용자에게 공지하는 절차와 사용자가 피드백을 인공지능 알고리즘 적용 및 운영 주체에 제공할 수 있는 절차를 평가해야 함

(2) 인공지능 알고리즘이 특정 서비스에 적용된 상태로 공교육 현장의 사용자에게 제공될 경우, 인공지능 알고리즘이 적용되고 있다는 사실이 특정 서비스를 통해서 사용자에게 공지되는 절차와 사용자가 피드백을 인공지능 알고리즘 적용 및 운영 주체 또는 특정 서비스를 제공하는 주체에게 제공할 수 있는 절차를 평가해야 함

마. 서울형 공교육 적용 인공지능 알고리즘 고위험군* 관리체제(안)

ㅇ 인공지능 알고리즘 위험성 평가 매트릭스와 영향평가 결과 고위험군으로 평가된 인공지능 알고리즘의 경우에는 별도의 관리 체제에 따라 관리함

*예) 학습자의 생체 정보를 활용한 자동화된 학습자 진단/평가 알고리즘

*예) 교원의 수업실행 및 학생관리 정보를 활용한 자동화된 교원 업무 능

력 진단/평가 알고리즘

*예) 행정직원의 업무 수행 정보를 활용한 자동화된 직원 업무 능력 진단/평가 알고리즘

ㅇ 공교육 적용 인공지능 알고리즘 고위험군 관리체제

(1) 고위험군 관리 위원회* 설치 운영

* 교원, 학생, 학부모, 행정직원 등 인공지능 알고리즘에 의한 자동화된 의사결정에 영향을 받을 가능성이 있는 이해관계자들과 전문가로 구성된 고위험군 관리 위원회 운영

(2) 고위험군 알고리즘의 공교육 현장 적용을 위한 시범 적용 프로세스 구축 및 적용 (테스트베드 적용, 6개월 이상의 모니터링, 위험성 평가 실행)

(3) 고위험군 알고리즘의 투명성/신뢰성/설명가능성 원칙 준수 여부 평가

(4) 고위험군 알고리즘의 기계학습에 사용되는 데이터세트 등록/공개*

* 위원회 공개, 전문가 그룹 공개, 일반 공개 등 단계별 공개 수준 결정

(5) 고위험군 기계학습에 사용되는 데이터세트 처리 과정/방법에 대한 정보 등록/공개*

(6) 자동화된 의사결정에 따라 발생할 가능성이 있는 차별 모니터링 및 재발 방지 방법 등록/공개*

(7) 고위험군 알고리즘이 발생시킬 가능성이 있는 위험성에 대해서, 일반인들도 수월하게 이해할 수 있는 언어로 작성/공개

4. 향후 추진 과제 제안

가. 공교육 인공지능 알고리즘 원칙(헌장), 영향평가 도구와 프로세스 등 개발

ㅇ 모든 학생과 교원에게 이로울 수 있는, 공교육 인공지능 알고리즘 원칙(헌장) 개발

- 앞으로 교육부/시도교육청 등의 공교육 기관에서는 인간에게 이로우며 동시에 신뢰할 수 있는 인공지능 알고리즘에 대한 기준을 구체적으로 제시하는 '신뢰가능 공교육 인공지능 원칙'을 고안하고 모든 학생과 교원이 인공지능의 혜택을 누릴 수 있는 기반을 조성해야 함

ㅇ 공교육 인공지능 알고리즘 영향평가(위험성 매트릭스 등) 도구와 프로세스 개발

- 다양한 공교육 영역에 적용될 가능성이 높은 인공지능 알고리즘이 해당 영역에서 야기할 가능성이 있는 위험성의 수준과 발생가능성을 평가할 수 있는 평가 도구와 평가 프로세스가 개발되어야 함

- 위험성의 수준에 따라, 인공지능 알고리즘을 관리할 수 있는 등급제 등이 방법을 고안해야 함

- 위험성의 수준에 따라, 시범 적용 등의 프로세스를 진행할 수 있는 테스트 베드* 여건을 고안해야 함

* 과학관 등에서 추진되고 있는 AI Lab 등의 체험공간을 테스트베드 공간으로 활용 가능

ㅇ 공교육 인공지능 알고리즘 조달 가이드라인 개발

- 공교육 기관에서 필요한 자원을 구매하기 위해 활용하는 조달 프로세스에 적용할 수 있는 조달 가이드라인이 개발되어 적용되어야 함

나. 공교육 인공지능 알고리즘 원칙의 실행을 위한 추진기구 설립

ㅇ 모든 학생과 교원에게 이로울 수 있는, 신뢰가능한 인공지능을 공교육에 적용하기 위한 원칙을 구체적인 정책으로 고안하고 실행할 수 있는 추진기구를 설립하여야 함

ㅇ (추진기구의 역할) 공교육에 적용되는 인공지능 알고리즘 원칙(헌장) 및 가이드라인 개발

- 공교육 인공지능 알고리즘 원칙(헌장) 개발
- 조달 가이드라인 등의 공교육 인공지능 가이드라인 개발
- 공교육 인공지능 알고리즘 영향 평가 방법 및 절차 개발

ㅇ (추진기구의 역할) 인공지능이 교육에 미칠 영향 평가 (공교육 적용 인공지능 알고리즘 영향평가) 체제 구축

- 인공지능을 공교육에 적용함에 따라 발생할 가능성이 있는 긍정적/부정적 영향을 평가하고, 그 결과에 따라 위험성을 차등 관리하는 추진기구의 역할이 필요함

ㅇ (추진기구의 역할에 따른 업무 흐름) 추진기구의 역할을 업무 흐름으로 구분해보면,

- 공교육 인공지능 알고리즘 원칙(헌장) 개발 및 발표
- 공교육 인공지능 알고리즘 영향 평가 도구(위험성 평가 매트릭스 등) 개발
- 공교육 인공지능 알고리즘 영향 평가 프로세스 구축
- 공교육 인공지능 알고리즘 영향 평가에 따른 위험성 등급 평가
- 위험성 등급에 따른 이후 처리 절차(테스트베드 상황에서의 시범 적용 등) 정의
- 영향 평가 사후 처리 결과를 공개하고 후속 과정 진행

ㅇ (추진기구의 역할:투명성 확보) 인공지능 알고리즘 투명성 확보(정보 공개와 참여) 체제 구축

- 공교육에 적용되는 인공지능 알고리즘의 신뢰가능성, 투명성, 설명가능성을 확보하기 위해 데이터의 측정/수집 -> 데이터의 전처리 -> 알고리즘 모형의 선택 -> 기계학습에 적용 -> 알고리즘 모형의 수정 및 업그레이드 -> 데이터 처리 결과의 출력 등에 이르는 인공지능 시스템의 생애주기 전체에 걸쳐 이해관계자의 참여가 가능한 평가 체제를 구축해야 함

- 기계학습 데이터셋에 대한 정보, 데이터 처리 과정에 대한 정보, 차별 방지 방법, 인간 감독 주체에 대한 정보, 위험성(개인차원, 사회차원)에 대한 정보를 공개하도록 하는 체제를 마련해야 함

○ (추진기구의 역할:이해관계자 역량 강화) 학생과 교원 등의 이해관계자의 정보주체 역량을 강화할 수 있는 프로그램의 운영

- 인공지능 시스템의 생애주기 전체에 참여하여 정보를 제공받고, 토론하고, 의사결정하는 과정에 참여하는 활동을 역량 강화 프로그램으로 활용

○ (추진기구의 역할:거버넌스 체제 구축) 공교육에 적용되는 인공지능 알고리즘의 공공성 확보를 위한 거버넌스 체제 구축

- 공교육에 적용되는 인공지능 알고리즘 개발 전문기관 (회사, 연구소 등), 관련 정책을 실행할 교원 (전문교원, 교원단체 등), 학생과 학부모 등이 참여하는 협의체와 인공지능 알고리즘이 도입되는 개별 정책의 실행과정과 결과를 관리 감독할 수 있는 거버넌스 체제를 구축해야 함

****

저자들이 참여하고 있는 사단법인 정보인권연구소는 디지털 기술이 사회구석구석에 스며들고 있는 이 시대에 그리고 앞으로의 시대에, 보호되어야 할 것은 무엇이고 변화되어야 할 것은 무엇인지, 결국 우리는 무엇을 추구해야하는지를 '인권'의 관점에서 탐구하는 연구집단입니다.

인공지능이 교육현장에 적용되기 시작한 현재, 인공지능을 (기술적, 개념적, 실천적, 윤리적으로) 어떻게 다루어야 할지를 궁리하고 있는 모든 분들, 특히 인공지능에게 전달되어 분석되고 처리될 가능성이 높아 보이는 학생의 개인정보, 교육과정이라는 이름으로 불린 '인류의 지식 체계'를 어떻게 다루어야 할지에 대해서 궁리하는 분들에게 이 보고서를 드립니다.

<인권과 기술> 토론서[59]

Human Rights and Technology : DISCUSSION PAPER

●

2019. 12.
호주 국가인권위원회

A. 도입 및 체계

❍ 제안 1: 호주 정부는 신흥기술에 대한 국가 전략을 마련해야 한다. 이 국가 전략은 다음과 같아야 한다.

(a) 국가적 목표를 책임감 있는 혁신을 촉진하고 인권을 보호하는 데 둔다.

(b) 인공지능에 대한 국가의 리더십 확보에 우선순위를 할당하고 자원을 제공한다.

(c) 법률, 공동 규제 및 자율 규제를 비롯하여 효과적인 규제를 촉진한다.

(d) 정부, 산업계 및 시민 사회를 교육 훈련할 수 있는 자원을 제공한다.

59) 2019년 12월 호주국가인권위원회는 <인권과 기술> 토론서에서 인공지능 등 신흥기술과 관련한 인권 문제를 살펴 보고 호주정부에 대한 30개의 제안 및 9개 질의 항목을 발표함. 이 부록은 그중 일부인 호주 정부에 대한 제안 및 질의 항목을 번역소개함.
<https://tech.humanrights.gov.au/sites/default/files/2019-12/TechRights_2019_DiscussionPaper.pdf?mc_cid=4d27f3ef7c&mc_eid=144114f10c>.

❍ 제안 2: 호주 정부는 신흥기술 윤리 체제에 대하여 다음과 같은 조사를 적절한 독립기구에 의뢰해야 한다.

(a) 현행 윤리 체제의 인권 보호 및 증진 효과를 평가한다.

(b) 유사 윤리 체계를 통합 또는 조화시키고 특정 기준을 충족하는 윤리 체계에 대해 특별법적 지위를 부여하는 등 윤리 체계 운용을 개선할 방안을 찾아본다.

B. 인공지능

〈질의 A〉 위원회가 제안한 '인공지능 정보 기반 의사결정'(AI-informed decision making)의 정의에는 다음의 두 가지 요소가 있다. 즉, 개인에게 법적 또는 이와 유사하게 중대한 영향을 미치는 의사결정이 있어야 하며, 또한 인공지능이 의사결정 과정을 실질적으로 지원했어야 한다. 인권 보호 및 기타 핵심 목표를 보호하기 위한 규제 목적상 '인공지능 정보 기반 의사결정'에 대한 위원회의 정의가 적절한가?

❍ 제안 3: 호주 정부는 호주법률개혁위원회(Australian Law Reform Commission)를 소집하여 인공지능 정보 기반 의사결정의 책무성에 대한 조사를 실시해야 한다. 이 조사는 다음을 위해 필요한 개혁이나 기타 변경을 고려해야 한다.

(a) 법률주의 및 법치주의 원칙을 보호하다

(b) 평등 및 비차별과 같은 인권을 증진한다.

❍ 제안 4: 호주 정부는 심각한 사생활 침해 행위에 대한 소송 청구원인(cause of action)을 법정화해야 한다.

❍ 제안 5: 호주 정부는 개인의 권리에 대해 법적 또는 이와 유사하게 중대한 영향을 미치는 의사결정에 인공지능이 실질적으로 사용된 경우 그 사람에게 정보를 제공하도록 하는 법안을 마련해야 한다.

❍ 제안 6: 호주 정부가 인공지능 정보 기반 의사결정 시스템을 도입하려고 계획할 경우, 다음과 같이 해야 한다.

(a) 인공지능 사용에 대한 비용 편익 분석을 수행하며 특히 인권 보호 및 책무 보장과 관련하여 살펴 본다.

(b) 가장 영향을 받을 가능성이 높은 사람들에 초점을 맞춘 공청회를 개최한다.

(c) 법률에 명시되고 적절한 인권 보호가 이루어진 경우에만 시스템을 도입한다.

❍ 제안 7: 호주 정부는 인공지능 정보 기반 의사결정의 설명가능성에 대한 법안을 마련해야 한다. 이 법안은, 인공지능을 사용하지 않은 의사결정에 대해 설명을 요구할 자격이 있었던 사람이라면 [인공지능 정보 기반 의사결정에 대하여] 다음을 요구할 수 있어야 한다는 점을 명시해야 한다.

(a) 인공지능 정보 기반 의사결정에 대하여 일반인에게 이해가능한 비기술적 설명

(b) 인공지능 정보 기반 의사결정에 대하여 관련 기술 전문성을 보유한 사람이 평가하고 검증할 수 있는 기술적 설명

각각의 설명에는 의사결정 사유를 포함해서, 개인 또는 관련 기술 전문가가 의사결정의 기반을 이해하고 이의를 제기해야 할 근거를 이해할 수 있어야 한다.

❍ 제안 8 : 인공지능 정보 기반 의사결정 시스템이 그 결정에 대해 합리적인 설명을 제공하지 않는 경우, 의사결정이 개인의 인권을 침해할 수 있는 상황에서 해당 시스템을 도입해서는 안 된다.

〈질의 B〉 어떤 사람이 인공지능 정보 기반 의사결정에 책임이 있고 그 사람이 그 결정에 대해 합리적인 설명을 제공하지 않는 경우, 그 결정이 합법적으로 내려지지 않았다는 반증허용추정(rebuttable presumption, 법정에서 설득력 있는 증거가 충분하게 있으면 뒤집어질 수 있는 추정 사실-역주)을 호주 법률에 도입해야 하는가?

❍ 제안 9 : 새로 설립된 〈자동화된 의사결정 및 사회에 대한 호주 연구거점 위원회(Australian Research Council Centre of Excellence for Automated Decision-Making and Society)〉 등 전문기술센터들은 개인에게 합리적인 설명을 제공하기 위한 인공지능 정보 기반 의사결정 시스템의 설계 방법에 대한 연구에 우선순위를 할당해야 한다.

❍ 제안 10: 호주 정부는 인공지능 정보 기반 의사결정 시스템을 도입하는 법인이 이 시스템의 사용에 대해 법적으로 책임을 져야 한다는 반증허용추정을 담은 법안을 마련해야 한다.

〈질의 C〉 알고리즘 등 인공지능 정보 기반 의사결정 시스템에 사용된 기술 정보에 더 수월한 접근을 제공함으로써, 인공지능 정보 기반 의사결정 시스템의 합법성을 보다 용이하게 평가할 수 있도록 호주 법을 개혁할 필요가 있는가?

〈질의 D〉 호주 법은 인공지능 정보 기반 의사결정 과정에서 인간 의사결정자의 개입을 어떻게 요구하거나 장려해야 하는가?

❍ 제안 11: 호주 정부는 적절한 법적 체제가 마련될 때까지 개인에게 법적 또는 이와 유사하게 중대한 영향을 미치는 의사결정에서 얼굴인식 기술의 사용에 대해 법적 유예(moratorium)를 도입해야 한다. 이 법적 체제는 강력한 인권 보호를 포함해야 하며 호주 국가인권위원회 및 호주 개인정보보호위원회 등 전문 기관과 협의하여 추진되어야 한다.

❍ 제안 12: 인공지능 정보 기반 의사결정과 관련하여 호주에 적용되는 모든 표준은 인권 준수에 대한 지침을 포함해야 한다.

❍ 제안 13: 호주 정부는 인공지능 정보 기반 의사결정 상황에서 '인권 중심 설계(human rights by design)' 개념을 발전시키기 위한 대책팀을 설립하고 호주에서 이를 어떻게 가장 잘 이행할 것인가를 검토해야 한다. 자율적 또는 법적으로 집행 가능한 인증 제도를 고려해야 한다. 대책팀은 이 분야에서 공공 및 민간 계획들에 대한 조정을 촉진하고, 인공지능 정보 기반 의사결정에 의해 인권이 중대하게 영향을 받을 가능성이 있는 사람들을 포함해 폭넓게 협의해야 한다.

❍ 제안 14: 호주 정부는 규제 당국, 산업계 및 시민사회 기관들과 협의하여 인공지능 정보 기반 의사결정을 위한 인권영향평가 도구와 그 사용에 대한 관련 지침을 개발해야 한다. 호주 정부가 보증한 '윤리적 인공지능을 위한 툴킷'들, 법률 체계 또는 지침들은 모두 인권영향평가를 명시적으로 포함해야 한다.

<질의 E> 제안 14의 인권영향평가 도구와 관련하여

(a) 언제 어떻게 도입해야 할 것인가?

(b) 인권영향평가 완수가 의무화되어야 할 것인가, 아니면 다른 방법으로 장려할 것인가?

(c) 평가에서 인권에 미치는 영향이 높은 위험도를 나타낸 경우 그 결과를 어떻게 처리할 것인가?

(d) 해외에서 개발된 인공지능 정보 기반 의사결정 시스템에 인권영향평가를 어떻게 적용해야 하는가?

❍ 제안 15: 호주 정부는 인공지능 정보 기반 의사결정 시스템의 인권 준수를 시험하기 위해 규제 샌드박스의 추진을 고려해야 한다.

<질의 F> 인공지능 정보 기반 의사결정 시스템의 인권 준수를 시험하기 위한 규제 샌드박스의 핵심 요소는 어떠해야 할 것인가? 특히,

(a) 참여 자격 기준이나 적용되는 시스템 유형 등 규제 샌드박스 운영 범위는 어디까지인가?

(b) 규제의 어떤 영역이 포함되어야 하는가? 예를 들어 인권 만인가 아니면 다른 영역도 포함할 것인가?

(c) 규제 샌드박스에서 제품을 승인하기 전에 어떤 통제 또는 기준을 적용해야 하는가?

(d) 어떤 보호나 인센티브로 참여를 촉진할 것인가?

(e) 어떤 기관(들)이 규제 샌드박스를 시행할 것인가?

(f) 규제 샌드박스가 관련 규제 당국 및 감독 기관, 시민 사회 및 산업계의 전문 지식을 어떻게 활용할 수 있는가?

(g) 경합하는 요구들(예: 투명성 대 영업비밀보호)의 균형을 어떻게 맞춰야 하는가?

(h) 규제 샌드박스를 어떻게 평가해야 하는가?

❍ 제안 16: 신흥기술에 대한 국가 전략(제안 1 참조)은 인공지능과 인권에 대한 교육을 포함해야 한다. 이는 일반 대중, 또는 인공지능 데이터포인트에 의존하는 의사결정자나 인공지능 정보 기반 의사결정 시스템을 설계하고 개발하는 직업군을 비롯하여 보다 전문적인 지식을 필요로 하는 사람 등, 특정한 기술과 지식에 대한 지역사회 여러 부문의 요구에 맞춘 교육 훈련이 포함되어야 한다.

❍ 제안 17: 호주 정부는 다음을 위해 새로운 기관 또는 기존 기관이 감독하는 포괄적인 검토를 수행해야 한다.

(a) 호주 정부의 의사결정 시 인공지능의 사용 여부 확인

(b) 인공지능 사용에 대한 비용편익 분석을 수행하며, 이때 인권 보호 및 책무성 보장을 특별히 살펴볼 것

(c) 인권영향평가를 비롯하여, 호주 정부가 인공지능을 사용하는 의사결정 시스템을 채택하기로 결정하는 절차의 개괄

(d) 의사결정으로 영향을 받는 사람에게 인공지능 사용에 대해 설명하는지 여부 확인, 가장 영향을 받을 가능성이 높은 사람들에 초점을 맞춘 공청회를 실시하는지 등 설명 방법에 대한 확인

(e) 의사결정에서 인공지능 사용에 대한 감시 및 평가 체계 검토

❍ 제안 18: 호주 정부 조달 규칙은 정부가 인공지능 정보 기반 의사결정 시스템을 조달할 경우 이 시스템에 적절한 인권 보호를 포함하도록 요구해야 한다.

C. 인공지능에 대한 국가의 리더십

❍ 제안 19: 호주 정부는 독립적인 법정 기구로 인공지능 안전위원회(AI Safety Commissioner)를 설립하여 호주 내 인공지능 개발과 이용에 있어 국가적으로 지도적인 역할을 담당하도록 해야 한다. 이 인공지능 안전위원회는 개인 및 지역사회의 피해를 방지하고 인권을 보호하고 증진하는데 주력해야 한다. 인공지능 안전위원회는 다음 역할을 수행해야 한다.

(a) 인공지능 개발 및 이용에 관한 기존 규제 기관 및 기타 기관의 역량을 구축한다.

(b) 인공지능 사용을 감시하고, 이 분야 정책 전문가의 원천이 된다.

(c) 그 조직 구조, 운영 및 입법 권한에서 독립적으로 활동한다.

(d) 호주 정부가 전적으로 또는 대부분의 재원을 적절하게 조달한다.

(e) 다양한 전문 지식과 관점을 활용한다.

(f) 우선순위를 할당하고 자기 업무를 구체화해야 할 당면 과제를 결정한다.

D. 접근 가능한 기술

❍ 제안 20 : 연방, 주, 지역 및 지방 정부는 현행 WCAG 2.1 및 호주 표준 EN 301 549 및 후속 표준 등 공인된 접근성 표준을 준수하는 디지털 기술을 사용해야 한다. 이를 위해 모든 호주 정부는 다음을 이행해야 한다.

(a) 위 접근성 표준을 충족하는 방법으로 디지털 기술을 사용하는 상품, 서비스 및 시설의 조달을 촉진하는, ‘접근 가능한 조달 정책’을 채택한다. 이러한 정책은 또한 정부 조달에 있어 자사 활동에서 접근성 표준을 구현하는 기업을 선호하게 할 것이다.

(b) 쉬운 영어 버전 및 인간 고객 지원 등 접근 가능한 통신 서비스의 가용성을 증진하는 정책을 개발한다.

❍ 제안 21: 호주 정부는 WCAG 2.1 및 호주 표준 EN 301 549와 같은 접근성 표준에 대한 산업계 준수 실태에 대해 조사를 실시해야 한다. 표준 준수에 따른 인센티브에는 과세, 보조금 및 조달, 연구 및 설계, 산업별 우수 관행 촉진과 관련한 변경이 있을 수 있다.

❍ 제안 22: 호주 정부는 1992년 방송법(Cth)을 개정하여 국가 방송 서비스, 상업 방송 서비스 및 가입 방송 서비스에 대해 다음을 의무화해야 한다.

(a) 각 채널에 대해 매주 14시간 이상(연도별 증가) 오디오 설명

(b) 연도별 기준으로 자막 컨텐츠의 주간 최저 시간 증가

❍ 제안 23: 호주 표준청은 장애인 및 기타 이해관계자들과 협의하여, 소비재에 수반하여 접근 가능한 정보, 지침 및 훈련 자료 제공을 보장하는 호주 표준 또는 기술 규격을 개발해야 한다.

❍ 제안 24: 전국광대역통신망은 재정적으로 취약한 장애인에게 할인된 광대역 도매요금을 제공하기 위한 경제 모델링을 실시해야 한다.

<질의 G> 디지털 기술 요금 관련 장애인에 대한 접근성의 장벽을 제거하기 위해 민간 부문이 취할 수 있는 다른 조치는 무엇인가?

❍ 제안 25: 호주정부연석회의 장애개혁위원회는 다음 업무를 수행해야 한다.

(a) 호주의 연방, 주 및 영토 정부가 디지털 기술을 이용한 정부 서비스의 개발 및 제공에 있어서 '인권 중심 설계'를 채택하고 촉진하는 것에 전념하는 절차를 주도하고, 이 목표 달성의 진척 상황을 감시한다.

(b) 장애인을 위한 디지털 및 기타 기술에 대한 접근성을 개선하기 위한 정책적 조치를 다음 국가 장애 전략의 우선순위로 포함한다.

❍ 제안 26: 3차교육 및 직업교육 제공자는 관련 학위 및 기타 과학, 기술, 공학 분야 과정에 '인권 중심 설계' 원칙을 포함해야 한다. 호주학술연구협의회는 적절한 지원을 통해 3차교육 및 직업교육 부문에서 이 목표를 가장 효과적이고 적절하게 달성할 수 있는 방법에 대한 협의를 진행해야 한다.

〈질문 H〉 ‘인권 중심 설계’에 대한 지침을 포함해야 하는 다른 3차 교육 과정이나 직업 교육 과정은 무엇인가?

❍ 제안 27: 공학, 과학, 기술 분야의 전문 인증 기구는 지속적인 전문성 개발의 일환으로 ‘인권 중심 설계’에 대한 의무 교육을 도입하는 것을 고려해야 한다.

❍ 제안 28: 호주 정부는 장애인에게 접근 가능한 기술의 교육, 훈련, 인증 및 역량 구축에 있어서 국가적인 개발 및 제공을 주도하기 위하여 조직적으로 위탁해야 한다.

❍ 제안 29: 호주 법무장관은 1992년 장애인차별금지법(Cth) 제31조에 따라 디지털 통신기술표준을 개발해야 한다. 이 새로운 표준을 개발함에 있어서 법무장관은 특히 장애인 및 기술 부문과 함께 폭넓게 협의해야 한다. 이 표준은 정보통신기술, 가상현실, 증강현실 등 디지털기술을 채용하는 곳을 비롯하여 주로 통신에 이용되는 공공 재화, 서비스, 설비의 제공에 적용되어야 한다.

〈질문 I〉 호주 정부가 1992년 장애인차별금지법(Cth)에 따라 디지털 기술에 대한 다른 유형의 표준들을 개발해야 하는가? 그렇다면, 이 표준들은 무엇을 포함해야 하는가? ▢

유럽연합 인공지능 백서[60]

WHITE PAPER On Artificial Intelligence

- A European approach to excellence and trust

●

2020. 2.
유럽연합 집행위원회

5. 신뢰 생태계 구축: AI 규제 프레임워크

다른 신기술과 마찬가지로 AI의 사용은 기회와 위험 모두를 가져온다. 시민들은 알고리즘 기반 의사결정의 정보 비대칭성에 직면하였을 때 자신의 권리와 안전을 방어하는 데 무력해지는 상태를 두려워하고 있으며, 기업들은 법적 불확실성을 우려하고 있다. AI가 시민의 안전을 보호하고 기본권을 누릴 수 있도록 기여할 수 있지만, 시민들은 AI가 의도치 않은 영향을 미치거나 심지어 악의적인 목적으로 사용될 수 있다고 우려한다. 이러한 우려는 해소될 필요가 있다. 투자와 기술력 부족 외에도 신뢰 부족이 AI의 광범위한 활용을 저해하는 주요 요인이다.

이것이 집행위가 2018년 4월 25일 AI 전략을 수립한 이유이다. 이 전략은 EU 전역의 연구, 혁신, AI 역량에 대한 투자 확충과 병행하여 사회경제적 측면을 다루고 있다. 집행위는 회원국들과 전략을 조정하기로 하고

60) 2020년 2월 유럽연합 집행위원회는 <인공지능 백서 - 수월성과 신뢰를 위한 유럽의 접근>에서 인공지능 규제 프레임워크를 발표함. 이 부록은 그중 일부인 5장과 6장의 내용을 번역소개함.
<https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf>.

통합 계획에 합의했다. 또한 집행위는 2019년 4월 고위전문가그룹을 발족시키고 &신뢰할 수 있는 AI 가이드라인&을 발간하였다.

집행위는 고위전문가그룹 가이드라인에서 확인한 7가지 주요 요구사항을 환영하는 내용의 공보(communication)를 발행했다. 이는 다음과 같다.

- 인적 개입 및 감독 (Human agency and oversight)
- 기술적 견고성 및 안전성
- 개인정보 보호 및 데이터 거버넌스
- 투명성
- 다양성, 차별금지 및 공정성
- 사회·환경적 복지
- 책무성

또한 이 가이드라인에는 기업의 실제 사용을 위한 평가 항목이 수록되어 있다. 2019년 하반기 동안 350개 이상의 기관이 이 평가 항목을 테스트하고 피드백을 보냈다. 고위전문가그룹은 이 피드백을 토대로 가이드라인을 개정하는 중이며 2020년 6월까지 이 작업을 마무리할 예정이다. 주요 피드백의 내용으로는, 가이드라인의 요구사항 다수가 이미 현행 법률 또는 규제 체제에 반영되어 있지만, 투명성/추적성/인적 감독에 관한 요구사항은 여러 경제 부문의 현행 법률에서 구체적으로 다루고 있지 않다는 것이었다.

고위전문가그룹의 구속력이 없는 일련의 가이드라인에 더하여, 그리고 의장의 정치적 방침에 의거하여, 유럽의 명확한 규제 프레임워크는 AI에 대한 소비자와 기업 간의 신뢰를 쌓고, 그에 따라 기술 활용의 속도를 높일 것이다. 이러한 규제 프레임워크는 이 분야에서 유럽의 혁신 역량과 경쟁력을 촉진하기 위한 다른 조치와 일관되어야 한다. 또한, 사회적, 환경적, 경제적으로 최적의 결과를 보장하고 EU 법률, 원칙 및 가치의 준수를 보장해야 한다. 이는 특히 법집행 및 사법 영역에서 AI를 응용하는 경우에서처럼, 시민의 권리에 가장 직접적인 영향을 미칠 수 있는 영역과 관련이 있다.

AI의 개발 및 도입 기관은 이미 기본권(예: 개인정보 보호, 프라이버시,

차별금지), 소비자 보호, 제품 안전 및 책임 규율에 관한 유럽 법률의 적용을 받고 있다. 소비자들은 제품이나 시스템이 AI에 의존하든 그렇지 않든 같은 수준의 안전과 권리의 존중을 기대한다. 그러나 AI의 일부 특징(예: 불투명성)은 이 법률들의 적용과 이행을 더 어렵게 할 수 있다. 이 때문에 현행 법률이 AI의 위험성을 해소할 수 있고 효과적으로 이행될 수 있는지, 법률의 개정이 필요한지 또는 새로운 입법이 필요한지 등을 따져 볼 필요가 있다.

AI가 빠르게 진화하고 있다는 사실을 감안하면, 규제 프레임워크는 장래의 개발에도 적용될 여지를 남겨야 한다. 규제 프레임워크에 대한 변경은 실현가능한 해결책이 존재하는 명확하게 확인된 문제로 제한되어야 한다.

회원국들은 현재 공통적인 유럽 프레임워크가 부재하다는 사실을 지적하고 있다. 독일 데이터윤리위원회는 5단계 위험 기반 규제 체제를 요구했다. 5단계는 가장 무해한 AI 시스템에 대한 무규제로부터 가장 위험한 AI 시스템에 대한 완전한 금지로 구성되어 있다. 덴마크는 최근 데이터 윤리 인증(Data Ethics Seal)에 대한 프로토타입을 출시했다. 몰타는 AI 자율인증제를 도입했다. EU가 EU 전체에 걸친 접근법을 제공하지 못한다면, 내부 시장에서 분열될 실질적인 위험이 있으며, 이는 AI에 대한 신뢰, 법적 확실성 및 시장 수용이라는 목표를 훼손할 것이다.

신뢰할 수 있는 AI에 대한 유럽의 건실한 규제 프레임워크는 모든 유럽 시민을 보호하고, AI의 장래 발전과 수용은 물론 마찰없는 내부 시장을 창출하여 유럽의 산업 기반을 강화하는 데 도움이 될 것이다.

A. 문제 정의

AI는 제품과 공정을 더 안전하게 만드는 등 많은 혜택을 줄 수 있지만, 위해를 끼칠 수도 있다. AI의 위해성은 유형적(생명 손상 등 개인의 안전과 건강, 재산 손실 등의 문제)이거나 무형적(사생활 침해, 표현의 자유 제약, 인간 존엄성, 고용 차별 등의 문제)일 수 있으며, 다양한 위험성과 관련될 수 있다. 규제 프레임워크는 잠재적 피해의 다양한 위험성, 특히 중대한 피해를 최소화하는 데 주력해야 한다.

AI의 이용과 관련된 주요한 위험성은 기본권(개인정보와 프라이버시 보호 및 차별금지 등)을 보호하기 위해, 또 안전과 책임 관련 문제에 적용하기 위해 설계된 규율의 적용과 관련이 있다.

개인정보와 프라이버시 보호 및 차별금지 등 기본권 측면에서 위험성

AI의 사용은 EU가 수립한 가치에 영향을 미칠 수 있으며, 표현의 자유, 집회의 자유, 인간의 존엄성, 성별・인종 또는 민족적 기원・종교 또는 신념・장애・연령 또는 성적 지향에 따른 차별금지를 포함하여, 개인정보 및 사생활 보호, 효과적인 사법적 구제와 공정한 재판을 받을 권리 등 기본권에 대한 침해로 이어질 수 있으며, 특정 영역에서는 소비자 보호를 침해할 수 있다. 이러한 위험성은 AI 시스템의 전체적인 설계 결함(인적 감독에 관한 문제 포함)에서 기인했을 수도 있고, 편향된 데이터를 교정하지 않고 사용하는 데서 기인했을 수 있다.(예: 시스템이 여성과 관련하여 차선적인 결과를 이끌어내는 남성의 데이터를 주되게, 또는 유일하게 사용하여 훈련된 경우)

AI는 이전에는 인간만이 할 수 있었던 많은 기능을 수행할 수 있다. 그 결과 시민과 법인은 AI 시스템에 의해 취해졌거나 또는 AI 시스템의 지원을 받아 취해진 조치와 결정에 점점 더 종속될 것이며, 이는 때때로 이해하기 어려울 수 있으며 필요한 경우에도 효과적으로 문제를 제기하는 것이 어려워질 수 있다. 게다가, AI가 사람들의 일상 습관을 추적하고 분석할 가능성이 높아졌다. 예를 들어, 각국 기관이나 여타 기구들이 대량 감시를 위해 EU 개인정보 보호 및 다른 규율들을 위반하며 AI를 사용하거나, 회사가 직원들의 행동을 관찰하기 위해 AI를 사용할 잠재적 위험성이 있다. 대량의 데이터를 분석하고 그 관계를 식별함으로써, AI는 사람에 대한 데이터를 재추적하고 탈익명화하는 데 사용될 수 있으며, 그 자체로는 개인정보를 포함하지 않는 데이터셋에 대해서도 개인정보 보호 측면에서 새로운 위험성을 유발할 수 있다. AI는 또한 온라인 사업자들이 사용자들에게 정보의 우선순위를 정해주고 내용을 규제할 때도 사용된다. 처리된 데이터, 애플리케이션의 설계 방식 및 인적 개입의 범위는 표현의 자유, 개인정보 보호, 프라이버시 및 정치적 자유에 영향을 미칠 수 있다.

특정 AI 알고리즘이 범죄 재범 예측에 악용될 경우, 여성 대 남성 또는 내국인 대 외국인에 대해 다른 재범 예측 가능성을 표시하면서 성별과 인종적 편향성을 드러낼 수 있다.

*출처: Tolan S., Miron M., Gomez E. and Castillo C. “Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia“, Best Paper Award, International Conference on AI and Law, 2019

얼굴 분석에 사용되는 특정 AI 프로그램의 경우, 피부색이 밝은 남성의 성별을 결정하는 데는 낮은 오류율을 보였지만 피부색이 어두운 여성의 성별을 결정하는 데는 높은 오류율을 보여주면서 성별과 인종적 편향성을 드러냈다.

*출처: Joy Buolamwini, Timnit Gebru; Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:77-91, 2018.

편견과 차별은 모든 사회경제 활동에 내재된 위험이다. 인간의 의사결정은 실수나 편견에 면역이 되어 있지 않다. 그러나 같은 편견이 AI에서 나타날 때, 인간 행동을 통제하는 사회통제 메커니즘 없이 많은 사람들에게 영향을 주고 차별할 수 있다는 점에서 훨씬 더 큰 영향을 미칠 수 있다. 이는 AI 시스템이 구동 중에 &학습&할 때도 발생할 수 있다. 그러한 경우, 설계 단계에서 결과를 예방하거나 예측할 수 없다면, 위험성은 시스템 본래의 설계적 결함에서 기인하는 것이 아니라 시스템이 대규모 데이터셋에서 식별하는 상관관계나 패턴에서 실제적인 영향을 받는다.

불투명성(&블랙박스 효과&), 복잡성, 예측 불가능성 및 부분적으로 자율적인 작동을 비롯해 여러 AI 기술의 특성들은, 기본권을 보호하기 위한 현행 EU 법률상 규율들을 준수하는지 검증하는 것을 어렵게 만들 수 있으며, 효과적으로 이행하는 것을 방해할 수도 있다. 집행 당국과 피해자들은 AI가 개입된 상태에서 해당 결정들이 어떻게 내려졌는지 검증하는 수단이 부족할 수 있고, 그 결과 관련 규정들이 준수되었는지 여부도 검증하기 어려울 수 있다. 개인과 법인은 그러한 결정들이 자신에게 부정적인 영향을 미칠 수 있는 상황에서 효과적인 사법 수단에 접근하는 데 어려움을 겪을 수 있다.

안전 및 책임 체제의 효과적인 기능 측면에서 위험성

AI 기술이 제품과 서비스에 내장되어 있을 때 사용자에게 안전상 새로운 위험성이 나타날 수 있다. 예를 들어, 물체 인식 기술의 결함으로 인해, 자율주행차는 도로에서 물체를 잘못 식별하여 부상과 물질적 손해가 수반되는 사고를 일으킬 수 있다. 기본권에 대한 위험성과 마찬가지로, 이러한 위험성은 AI 기술의 설계상 결함에 의해 발생될 수 있으며, 이는 데이터의 가용성 및 품질상의 문제 또는 기계 학습에서 기인한 기타 문제와 관련되어 있다. 이러한 위험성의 일부는 AI에 의존하는 제품과 서비스에 국한되지 않지만, AI의 사용은 이러한 위험성을 증가시키거나 악화시킬 수 있다.

이러한 위험을 다루는 명확한 안전 규정의 부족은, 관련 개인에 대한 위험성과 별도로 EU에서 AI와 관련된 제품을 마케팅하는 기업에게 법적 불확실성을 야기할 수 있다. 시장 감시 및 집행 당국은 자신들이 개입할 수 있는지 여부에 대해 불분명한 상황에 처할 수 있는데, 이는 이들이 시스템을 검사하기 위한 집행권을 부여받지 못했거나 적절한 기술적 역량을 보유하지 못했기 때문이다. 따라서 법적 불확실성은 전반적인 안전 수준을 떨어뜨리고 유럽 기업의 경쟁력을 떨어뜨릴 수 있다.

안전 위험이 현실화될 경우, 위에서 언급한 AI 기술의 특징과 명확한 요구사항의 부족으로 인해 AI 시스템이 개입하여 내려진 문제적일 수 있는 의사결정을 역추적하는 것이 어려워진다. 이는 결국 피해를 입은 사람이 현행 EU 및 국가 책임 법률에 따라 보상을 받는 것을 어렵게 할 수 있다.

> 제품 책임 지침에 따르면 제조업체는 제품 결함으로 인한 손상에 대해 책임을 진다. 다만 자율주행차 등 AI 기반 시스템의 경우, 제품의 하자와 발생한 피해 간에 인과관계가 있다는 점을 입증하기 어려울 수 있다. 또한 제품의 사이버 보안상의 취약점으로 인한 문제 등 특정 유형의 결함의 경우, 제품 책임 지침이 적용되는 방법과 정도 측면에서 약간의 불확실성이 있다.

따라서, 기본권과 관련하여 위에서 언급한 바 대로 AI 시스템에 의해 내려진 문제적일 수 있는 결정을 역추적하는 어려움은 안전 및 책임 관련

문제에도 동일하게 적용된다. 예를 들어, 위해를 입은 사람이 법정에서 사건을 구성하는 데 필요한 증거에 효과적으로 접근하지 못할 수 있으며, 전통적인 기술로 인해 손해를 입은 상황에 비해 효과적으로 시정될 가능성이 더 적을 수 있다. AI의 사용이 더욱 확산되면 이러한 위험성이 증가할 것이다.

B. AI 관련 현행 EU 법률 체계의 조정 가능성

광범위한 현행 EU 제품 안전 및 책임 법률은 부문별 규율을 포함하고 회원국 국내법에 의해 더욱 보완되는데, 새로 부상하는 다수의 AI 애플리케이션과도 관련이 있으며 잠정적으로 적용할 수 있다.

기본권과 소비자 권리의 보호와 관련하여, EU의 법률 체계는 인종 평등 지침, 고용 및 직업에서 동등한 처우에 관한 지침, 재화 및 서비스에 대한 접근과 고용에서 남녀 동등 처우에 관한 지침, 일련의 소비자 보호 규정들은 물론, GDPR로 대표되는 개인정보 보호 및 프라이버시에 대한 규정들 및 법집행기관 개인정보 보호 지침 등 기타 부문별 개인정보 보호 법률을 아우른다. 또 2025년부터는 유럽접근성법에 규정된 상품 및 서비스에 대한 접근성 요구사항에 관한 규율들이 적용된다. 또한, 금융 서비스, 이주, 혹은 온라인사업자의 책임에 대한 분야를 비롯해 다른 EU 법률을 시행할 때 기본권을 존중할 필요가 있다.

EU 법률은 AI의 개입 여부와 관계없이 원칙적으로 충분히 적용 가능한 상태이지만, 법률이 AI 시스템이 창출하는 위험성을 해소하기 위해 적절히 시행될 수 있는지, 또는 특정 법률 수단에 대한 조정이 필요한지 여부에 대해 평가하는 것이 중요하다.

예를 들어, 경제 행위자는 소비자를 보호하는 현행 규정에 대한 AI의 준수 문제에 대해 전적으로 책임을 지며, 알고리즘이 현행 규정을 위반하여 소비자 행동을 부당하게 이용하는 것은 허용되지 않고 위반 행위도 그에 준하여 처벌되어야 한다.

집행위는 다음과 같은 위험과 상황에 대처하기 위해 법률 체계를 개선할 수 있다는 의견이다.

- 현행 EU 법률 및 회원국 국내법의 효과적인 적용 및 집행: AI의 주요 특성은 EU 법률 및 회원국 국내법의 적절한 적용과 시행을 보장하는 데 있어 문제를 야기한다. 투명성 부족(AI의 불투명성) 때문에 기본권 보호, 속성책임, 배상청구 조건 충족 등 법 위반 가능성을 확인하고 입증하기 어렵다. 따라서 효과적인 적용과 집행을 보장하기 위해서는, 예를 들어 본 백서에 수반되는 보고서에 추가적으로 설명된 책임에 관한 분야를 비롯해 특정 분야의 현행 법률을 조정하거나 명확히 할 필요가 있을 수 있다.

- 현행 EU 법률의 범위 제한: EU 제품안전법의 핵심적인 초점은 제품을 시장에 출시하는 경우에 대한 것이다. EU 제품안전법에서 소프트웨어는 최종 제품의 일부인 경우 관련 제품안전 규칙을 준수해야 하지만, 독립형 소프트웨어가 명시적 규칙을 가진 일부 부문 바깥에서 EU 제품안전법의 적용을 받는지 여부는 열린 질문이다. 현재 시행 중인 일반 EU 안전법은 서비스가 아닌 상품에 적용되며, 따라서 원칙적으로 AI 기술(예: 보건의료 서비스, 금융 서비스, 운송 서비스)에 기반한 서비스에는 적용되지 않는다.

- AI 시스템 기능 변경: AI를 비롯해 소프트웨어를 제품으로 통합하면 그 생애주기 동안 해당 제품과 시스템의 기능을 수정할 수 있다. 이것은 특히 빈번한 소프트웨어 업데이트가 필요하거나 기계 학습에 의존하는 시스템의 경우에 해당된다. 이러한 특징들은 시스템을 시장에 내놓았을 때 없었던 새로운 위험을 야기할 수 있다. 이러한 위험은 시장에 출시했을 때 나타나는 안전 위험에 주로 초점을 맞춘 현행 법률에서 적절히 다루고 있지 않다.

- 공급망에서 서로 다른 경제 행위자 간의 책임 배분에 관한 불확실성: 일반적으로 제품 안전성에 관한 EU 법률은 시장에 배치된 제품의 생산자에게 책임을 배분하는데, 이는 AI 시스템 등 모든 구성 요소를 포함한다. 그러나 예를 들어 생산자가 아닌 측에서 제품을 시장에 내놓은 후 AI가 추가되면 이 규칙이 불분명해질 수 있다. 또한, EU 제품책임법은 생산자의 책임을 규정하고, 공급망에서 다른 사람의 책임을 통제하는 것은 국내 책임법의 역할로 남겨두었다.

- 안전 개념의 변화: 제품과 서비스에서 AI를 사용하는 것은 EU 법률이

현재 명시적으로 다루지 않는 위험을 야기할 수 있다. 이러한 위험들은 사이버 위협, 개인 보안 위험(예: 가전제품과 같은 AI의 새로운 애플리케이션과 연계될 경우), 연결 손실에 따른 위험 등과 연결될 수 있다. 이러한 위험은 제품을 시장에 출시할 때 나타나거나 제품을 사용하는 동안 소프트웨어 업데이트 또는 자가 학습의 결과로 발생할 수 있다. AI 위협 요소 평가에 EU 사이버보안청(ENISA)의 경험을 활용하는 등, EU는 AI 애플리케이션과 연계된 잠재적 위험에 대해 증거 기반을 강화할 수 있는 관할 수단을 최대한 활용해야 한다.

앞에서 지적했듯이, 몇몇 회원국들은 이미 AI로 인해 야기된 난제를 해결하기 위해 국내 입법이라는 선택을 모색하고 있다. 이는 단일 시장이 분열될 수 있는 위험을 높인다. 국가간 규정이 엇갈리면 단일 시장에서 AI 시스템을 판매・운용하려는 기업에 장벽이 생길 가능성이 높다. EU 수준에서 공통적인 접근방식을 보장하면 유럽 기업들이 단일 시장에 대한 원활한 접근으로부터 이익을 얻고 글로벌 시장에서의 경쟁력을 지원받을 수 있을 것이다.

인공지능, 사물인터넷, 로봇공학이 갖는 안전과 책임의 의미에 관한 보고서

본 백서와 함께 제공되는 보고서는 관련 법적 체제를 분석하였다. 보고서는 AI 시스템 및 기타 디지털 기술에 의해 야기되는 특정 위험에 대하여 이 체제들을 적용할 때 나타나는 불확실성을 확인하였다.

보고서는 현행 제품안전법이 이미 제품 사용에 따라 제품에서 발생하는 모든 종류의 위험으로부터 안전을 보호하는 확장된 개념을 지지하고 있다고 결론내렸다. 그러나 새로 부상하는 디지털 기술이 드러내는 새로운 위험을 명시적으로 다루는 조항이 도입되면 보다 법적 확실성을 제공할 수 있을 것이다.

▸생애주기 동안 특정 AI 시스템의 자율적 작동은 안전에 영향을 미치는 중요한 제품 변경을 수반할 수 있으며, 이는 새로운 위험성 평가를 필요로 할 수 있다. 또한 제품 설계서부터 AI 제품 및 시스템의 생애주기 전체에 대한 안전장치로서 인적 감독이 필요할 수 있다.

▸생산자에 대한 명시적 의무는 적절할시 사용자의 정신적 안전 위험과 관련해서도 고려될 수 있다(예: 휴머노이드 로봇과의 협업).

▸조합 제품안전법은 AI 제품 및 시스템 사용 전반에 걸쳐 데이터 품질을 유지할수 있는 메커니즘뿐만 아니라 설계 단계에서 결함 있는 데이터의

안전성 위협을 해결하는 특정 요구사항을 규정할 수 있다.

▸알고리즘에 기반한 시스템의 불투명성은 투명성 요구사항을 통해 해결할 수 있다.

독립형 소프트웨어가 있는 그대로 시장에 배치되었거나 시장에 배치된 후 제품에 다운로드되었는데 안전성에 영향을 미치는 경우, 현행 규칙을 조정하고 명확히 할 필요가 있을 수 있다.

▸신기술 공급망의 복잡성이 증가함에 따라, 공급망 내 경제 행위자와 사용자 간의 협력을 특별히 요청하는 조항은 법적 확실성을 제공할 수 있다.

AI, 사물인터넷, 로봇공학과 같은 새로운 디지털 기술의 특성은 책임 프레임워크의 측면에서 문제를 야기할 수 있고 그 효과를 감소시킬 수 있다. 이러한 특성들 중 일부는 사람에게 돌아가 그 피해를 추적하는 것을 어렵게 만들 수 있으며, 이는 대부분의 국내 규율에 따라 결함에 기반해 청구를 제기할 때 필요한 것들이다. 이는 피해자 비용을 상당히 증가시킬 수 있으며, 생산자 외 타인에 대한 책임 청구가 이루어지거나 입증하기 어려울 수 있다는 것을 의미한다.

▸AI 시스템의 개입으로 피해를 입은 사람은 다른 기술로 피해를 입은 사람과 동일한 수준의 보호를 누릴 필요가 있는 한편으로, 기술혁신은 계속 발전할 수 있도록 해야 한다.

▸제품 책임 지침에 대한 수정 가능성 및 국내 책임법들의 추가적인 표적형 조정 가능성을 비롯하여, 이상의 목적을 보장하기 위한 모든 선택지들을 신중하게 평가해야 한다. 예를 들어, 집행위는 AI 애플리케이션 운영으로 인한 피해에 대해 국내책임법에서 요구하는 입증책임을 적용함으로써 복잡성의 결과를 완화하는 것이 필요한지 여부와 그 정도에 대해 생각해보고 있다.

집행위는 위와 같은 논의로부터, 현행 법률에 대한 조정 가능성과 별도로, EU의 법적 체제가 현재 및 예상되는 기술과 상업적 발전에 부합하도록 AI에 특화된 새로운 법률이 필요할 수 있다고 결론짓는다.

C. 미래 EU 규제 프레임워크의 범위

향후 AI에 대한 구체적인 규제 프레임워크의 핵심 쟁점은 적용 범위를 결정하는 것이다. 이 규제 프레임워크는 AI에 의존하는 제품과 서비스에 적용된다는 것이 작업 가설이다. 따라서 이 백서의 목적뿐만 아니라 미래의 정책 수립을 위해 AI가 명확하게 정의되어야 한다.

집행위는 "유럽을 위한 AI"에 관한 공보에서 AI에 대한 첫 정의를 내렸다. 이 정의는 고위전문가그룹에 의해 더욱 다듬어졌다.

어떤 새로운 법적 수단에서든, AI의 정의는 필요한 법적 확실성을 제공할 수 있을 만큼 정확하면서 기술적 진보를 수용할 수 있을 만큼 충분히 유연해야 할 것이다.

> 예를 들어 자율주행에서 알고리즘은, 차량이 특정 목적지에 도달하기 위해 어떤 방향으로 어떤 가속 및 속도를 취해야 하는지 도출하기 위해 자동차의 데이터(속도, 엔진 소모, 충격 흡수장치 등)와 차량 주변 모든 환경을 스캔하는 센서의 데이터(도로, 신호, 다른 차량, 보행자 등)를 실시간으로 사용한다. 관찰된 데이터를 바탕으로 알고리즘은 도로의 상황과 다른 운전자의 행동을 포함한 외부 조건에 맞게 조정하여 가장 편안하고 안전한 운행을 끌어낸다.

본 백서의 취지는 물론 정책에 대한 향후 논의 측면에서 AI를 구성하는 주요 요소를 명확히 하는 것이 중요해 보이는데, 이는 「데이터」와 「알고리즘」이다. AI는 하드웨어에 통합될 수 있다. AI의 하위 집합을 구성하는 머신러닝 기법의 경우, 주어진 목표를 달성하는 데 필요한 조치를 결정하기 위해 데이터셋을 기반으로 특정 패턴을 추론하도록 알고리즘을 훈련한다. 알고리즘은 사용되는 동안 계속 학습할 수 있다. AI 기반 제품은 사전에 정해진 지침을 따르지 않고 자율적으로 환경을 인식하고 작동할 수 있지만, 이들의 작동은 대부분 개발자들에 의해 정의되고 제약받는다. 인간이 목표를 결정하고 프로그래밍하고, AI 시스템은 이를 최적화해야 한다.

EU는 특히 소비자를 보호하고 불공정한 상업 관행에 대응하며 개인정보와 사생활을 보호하기 위한 엄격한 법률 체계를 시행하고 있다. 또한

역내 조약에는 특정 부문(예: 의료, 운송)에 대한 특별 규정들이 포함되어 있다. 디지털 변환과 AI 사용을 반영하기 위해 이들 EU 법률 체계에 어느 정도 업데이트가 필요할 수 있지만, 현행 조항들은 AI에 대해서도 계속하여 적용될 것이다(섹션 B 참조). 따라서 수평적 또는 부문적 현행 법률(예: 의료기기 관련 법률, 운송 시스템 관련 법률)에서 이미 다루고 있는 측면은 계속하여 이들 법률이 적용될 것이다.

원칙적으로 AI에 대한 새로운 규제 프레임워크는 목표를 달성하는 데 효과적이면서 중소기업에 부담이 될 만큼 지나치게 지시적이지 않아야 한다. 이 균형을 맞추기 위해, 집행위는 위험 기반 접근법을 따라야 한다고 생각한다.

규제 개입이 비례적일 수 있으려면 위험 기반 접근법이 중요하다. 다만, 서로 다른 AI 애플리케이션을 구별할 수 있는 명확한 기준이 필요하고 이로써 이들이 '고위험'인지 여부를 결정할 수 있어야 한다. 고위험 AI 애플리케이션이 무엇인지에 대한 결정은 명확하고 쉽게 이해할 수 있어야 하며 모든 관련 당사자에게 적용할 수 있어야 한다. 그러나 설령 AI 애플리케이션이 고위험군으로 분류되지 않더라도 여전히 전적으로 현행 EU 규정의 적용을 받는다.

집행위는 AI 애플리케이션이 사용되는 분야와 예상되는 용도 모두에서 안전, 소비자 권리 및 기본권의 보호 측면에서 상당한 위험을 수반하는지 여부를 고려하여, 문제가 있는 경우 해당 AI 애플리케이션을 일반적으로 고위험으로 간주해야 한다는 의견이다.

구체적으로는 AI 애플리케이션이 다음의 두 가지 누적적 기준을 충족할 경우 고위험으로 간주되어야 한다.

- 첫째로, 일반적으로 수행되는 활동의 특성을 고려할 때 상당한 위험이 발생할 것으로 예상되는 분야에 해당 AI 애플리케이션이 배치되는 경우. 이 첫 번째 기준은 일반적으로 말해서 위험이 가장 발생할 가능성이 높다고 판단되는 영역을 대상으로 규제가 개입하도록 한다. 해당 부문은 새 규제 프레임워크에서 구체적이고 빠짐없이 열거되어야 한다. 예를 들어, 의료, 운송, 에너지 및 일부 공공 부문이 이에 해당할 것이다. 열거 목록은 주기적으로 검토되어 실무에서 관련 개발이 기능상 필요한 경우 수정되어야 한다.

● 둘째, 해당 분야 문제의 AI 애플리케이션이 추가적으로 상당한 위험이 발생할 가능성이 높은 방식으로 사용되는 경우. 이 두 번째 기준은 선택된 분야에서 사용되는 모든 AI가 반드시 상당한 위험을 수반하는 것은 아니라는 사실을 반영한다. 예를 들어, 보건의료는 일반적으로 관련 부문일 수 있지만, 병원 예약시스템의 결함은 일반적으로 법률적 개입을 정당화할 만큼 심각한 위험을 야기하지는 않을 것이다. 특정 용도의 위험 수준을 평가할 때는 관련 당사자들에 미치는 영향에 기초할 수 있다. 예를 들어, 개인이나 기업의 권리에 대해 법적인 영향 또는 유사하게 상당한 영향을 미치는 AI 애플리케이션을 사용하는 경우, 부상, 사망 또는 상당한 유무형적 손상을 초래하는 AI 애플리케이션을 사용한 경우, 개인이나 법인이 합리적으로 피할 수 없는 효과를 낳는 AI 애플리케이션을 사용하는 경우 등이 있다.

두 가지 누적 기준을 적용하면 규제 프레임워크의 적용범위의 대상이 분명하고 법적 확실성을 보장할 수 있다. AI에 관한 새로운 규제 프레임워크에 포함된 의무적 요구사항(아래 섹션 D 참조)은 원칙적으로 이 두 가지 누적 기준에 따라 고위험으로 확인된 애플리케이션에만 적용된다.

전술한 사항에도 불구하고, 그 위험성으로 인해 특정 목적의 AI 애플리케이션의 사용을 위험성이 높은 것으로 간주하는 예외적인 사례도 있을 수 있다. 이 경우 분야를 불문하고 다음 요구사항이 여전히 적용될 것이다. 대략적으로 다음과 같은 특별한 경우를 생각해볼 수 있다.

● 개인과 EU 조약에서 고용 평등의 중요성에 비추어 볼 때, 채용 과정과 근로자의 권리에 영향을 미치는 상황에서 AI 애플리케이션의 사용은 항상 “고위험“으로 간주될 수 있으며, 따라서 아래 요구사항이 항상 적용될 것이다. 나아가 소비자 권리에 영향을 미치는 특정 애플리케이션의 경우도 고려될 수 있다.

● 원격 생체인식 및 기타 침입 감시 기술 목적으로 AI 애플리케이션을 사용하는 것은 항상 “고위험“으로 간주되며, 따라서 아래 요구사항이 항상 적용된다.

D. 요구사항의 유형

향후 AI에 대한 규제 프레임워크를 설계할 때 관련 주체에게 부과할 의무적인 법적 요구사항의 유형을 결정해야 할 것이다. 이러한 요구사항들은 표준을 통해 보다 구체적으로 명시될 수 있다. 위의 섹션 C에서 언급한 바와 같이, 현행 법률의 규제에 더하여, 이들 요구사항들은 고위험 AI 애플리케이션에만 적용되므로 규제 개입의 대상을 집중시키고 비례적인 규제를 보장할 수 있다.

고위전문가그룹 가이드라인 및 앞서 설명한 사항을 고려하여, 고위험 AI 애플리케이션의 요구사항은 다음과 같은 주요 기능으로 구성될 수 있다. 더 자세한 사항은 하위 절에서 설명한다.

- 훈련용 데이터
- 데이터 및 기록 보존
- 제공해야 할 정보
- 견고성 및 정확성
- 인적 감독
- 원격 생체인식 목적 사용 등 특정 AI 애플리케이션에 대한 특별 요구사항

법적 확실성을 보장하기 위해 이러한 요구사항은 추가적으로 구체화되어 이를 준수해야 하는 모든 주체에게 명확한 기준을 제공할 것이다.

a) 훈련용 데이터

EU의 가치와 규칙, 특히 EU 법률에서 도출되는 시민의 권리를 증진하고 강화하며 옹호하는 것이 그 어느 때보다 중요하다. 이러한 노력은 의심할 여지 없이 이 백서에서 검토 중이고 EU에서 출시되어 사용될 고위험 AI 애플리케이션에게도 적용된다.

앞서 논의한 것처럼 데이터가 없으면 AI도 없다. 많은 AI 시스템의 기능, 그리고 이들이 이끌어낼수 있는 조치와 결정들은 시스템이 훈련하는

데 사용된 데이터셋에 매우 많이 의존한다. 따라서 AI 시스템 훈련에 사용되는 데이터에 관한 한, EU의 가치와 규율들, 특히 안전 및 기본권 보호를 위한 현행 법률 규정들이 존중되도록 필요한 조치를 취해야 한다. AI 시스템 훈련에 사용되는 데이터셋과 관련해 다음과 같은 요구사항을 생각해볼 수 있다.

- AI 시스템에 기반한 제품이나 서비스의 후속 사용이 해당 EU 안전 규칙(현행 및 가능한 보완 규칙)에 규정된 기준을 충족하는 등, 안전에 대하여 합리적으로 보증하려는 목적의 요구사항. 예를 들어, AI 시스템이 위험한 상황을 방지하는데 필요한 모든 관련 시나리오를 포괄하고 충분히 광범위한 데이터셋에 기반해 훈련하도록 하는 요구사항
- 이러한 AI 시스템의 후속 사용이 금지된 차별을 수반하는 결과로 이어지지 않도록 합리적인 조치를 취하기 위한 요구사항. 이러한 요구사항은 특히 젠더, 민족성 및 기타 금지된 차별과 관련된 모든 차원이 해당 데이터셋에 적절히 반영되도록 충분히 대표성 있는 데이터셋을 사용해야 할 특정한 의무를 수반할 수 있다.
- AI 탑재 제품과 서비스를 사용하는 동안 사생활과 개인정보를 적절히 보호하기 위한 요구사항. 이 문제에서는 GDPR과 법집행 지침이 각각의 범위에 해당하는 문제에 대해서 규제한다.

b) 기록 및 데이터의 보존

많은 AI 시스템의 복잡성과 불투명성, 그리고 해당 분야 규칙의 준수를 효과적으로 검증하고 시행하는 것과 관련된 어려움을 비롯해 여러 요소들을 고려해 볼 때, 알고리즘 프로그래밍과 관련된 기록, 고위험 AI 시스템 훈련에 사용되는 데이터, 특정한 경우에 데이터 자체의 보존에 대한 요구사항이 필요하다. 이러한 요구사항들은 기본적으로 문제가 될 수 있는 AI 시스템 기반 조치나 결정을 추적하고 검증할 수 있도록 한다. 이는 감독 및 시행을 용이하게 할 뿐만 아니라, 관련 경제 행위자들이 그러한 규칙을 존중해야 할 필요성에 대해 초기 단계서부터 고려하도록 동기유발을 강화할 수 있다.

이 목적을 위해 규제 프레임워크는 다음을 지켜야 한다고 규정할 수 있

다.

- AI 시스템의 훈련 및 테스트에 사용된 데이터셋에 대한 정확한 기록 (주요 특성 및 데이터셋 선택 절차에 대한 서술 포함)
- 정당한 경우, 데이터셋 그 자체
- 시스템의 구축/테스트/검증에 사용된 프로그래밍 및 훈련 방법론에 대한 문서화(관련성이 있는 경우 안전성 및 금지된 차별을 초래할 수 있는 편향의 방지와 관련된 사항 포함)

관련 법률을 효과적으로 이행하기 위해 기록, 문서 및 관련된 경우 데이터셋까지 제한적이고 합리적인 기간 동안 보존해야 한다. 특히 관할 당국에서 시험이나 검사에 대한 요청이 있을 경우 이용 가능하도록 조치해야 한다. 필요한 경우 영업비밀과 같은 기밀정보가 보호될 수 있도록 협의해야 한다.

c) 정보제공

위의 지점 c)에서 논의된 기록 보존 요구사항을 넘어 투명성 또한 요구된다. 특히 AI의 책임 있는 사용을 촉진하고, 신뢰를 구축하고, 필요한 경우 시정을 용이하게 하는 등 추구하는 목표를 달성하기 위해서는 고위험 AI 시스템 사용에 대해 사전 예방적 방식으로 적절한 정보를 제공하는 것이 중요하다.

따라서 다음과 같은 요구사항을 고려할 수 있다.

- AI 시스템의 역량과 한계, 특히 시스템이 의도한 목적, 의도한 대로 기능할 것으로 기대할 수 있는 조건 및 특정 목적을 달성하는 데 예상되는 정확도 수준에 대해 명확한 정보가 제공되도록 보장한다. 이 정보는 특히 시스템의 출시자에게 중요하지만, 관할 당국 및 영향을 받는 당사자와도 관련이 있을 수 있다.
- 이와는 별도로, 시민들이 사람이 아니라 AI 시스템과 상호작용할 때 시민들에게 이를 명확히 알려야 한다. EU 개인정보 보호 법률은 이미 이러한 종류의 특정 규칙을 포함하고 있지만, 전술한 목적을 달성하기 위해 추가 요구사항이 요구될 수 있다. 그렇다면 불필요한 부담은 피해

야 한다. 따라서, 예를 들어, 시민들이 AI 시스템과 상호작용하고 있다는 것이 즉시 명백해지는 상황에서는 그러한 정보를 제공할 필요가 없다. 더 나아가 제공된 정보가 객관적이고 간결하며 쉽게 이해할 수 있어야 하는 것이 중요하다. 정보가 제공되는 방식은 특정 상황에 맞게 조정되어야 한다.

d) 견고성 및 정확성

AI 시스템, 그리고 확실히 고위험 AI 애플리케이션은 신뢰할 수 있을 만큼 기술적으로 견고하고 정확해야 한다. 즉, 그러한 시스템은 책임감 있는 방식으로 개발되어야 하며, 발생 가능한 위험에 대한 사전적이고 적절한 고려가 있어야 한다. 이들의 개발과 기능은 AI 시스템이 의도한 대로 확실하게 작동해야 한다. 위해가 발생할 위험을 최소화하기 위해 모든 합리적인 조치를 취해야 한다.

따라서 다음과 같은 요소를 고려할 수 있다.

- 모든 생애주기 단계에서 AI 시스템이 견고하고 정확하도록 보장하거나 최소 자기 정확성의 수준을 올바르게 반영하도록 보장하는 요구사항
- 결과를 재현할 수 있도록 보장하는 요구사항
- AI 시스템이 모든 생애주기 단계에서 오류 또는 불일치를 적절히 처리할 수 있도록 보장하는 요구사항
- AI 시스템이 공개적인 공격은 물론 데이터 또는 알고리즘 자체를 조작하려는 보다 교묘한 시도 모두에 대해 탄력성을 갖도록 보장하고, 그러한 경우 완화 조치를 취하도록 보장하는 요구사항

e) 인적 감독

인적 감독은 AI 시스템이 인간의 자율성을 훼손하거나 다른 부작용을 일으키지 않도록 보장하는 데 도움이 된다. 신뢰할 수 있고 윤리적이며 인간 중심적인 AI를 위한 목표는 고위험 AI 애플리케이션과 관련하여 인간의 적절한 관여를 보장해야만 달성할 수 있다.

본 백서에서 특별 법률 체제에 대해 고려하는 AI 애플리케이션은 모두 고위험으로 간주되지만, 인간 감독의 적절한 유형과 정도는 사례마다 다를 수 있다. 특히 시스템의 의도된 사용과 해당 사용이 당사자 시민과 법인에 미칠 수 있는 영향에 따라 달라져야 한다. 또한 AI 시스템이 개인정보를 처리할 때 GDPR이 확립한 법적 권리에 대한 침해도 없어야 한다. 예를 들어, 인적 감독에는 다음과 같은 경우가 포함될 수 있다.

- AI 시스템의 결과물은 사전에 사람에 의해 검토되고 검증되지 않은 경우 효력이 없음 (예: 사회보장급여 신청에 대한 거부는 사람에 한하여 할 수 있음)
- AI 시스템의 결과물은 즉시 효력이 발생하지만, 사후 인적 개입이 보장됨 (예: 신용 카드 신청에 대한 거부는 AI 시스템으로 처리할 수 있지만, 사후 인적 검토가 가능해야 함)
- AI 시스템에 대한 작동 중 모니터링 및 실시간 개입, 비활성화 기능 (예: 무인 자동차에서 사람이 자동차 운행이 안전하지 않다고 판단할 때 정지 버튼이나 절차를 사용할 수 있음)
- 설계 단계에서, AI 시스템에 작동 제약을 가함 (예: 무인 자동차는 가시성이 저하되어 센서 신뢰도가 낮아진 특정 조건에서 작동을 중지하거나 어떤 조건에서도 선행 차량과 일정 거리를 유지해야 함)

f) 원격 생체인식에 대한 구체적인 요구사항

예를 들어 공공장소에 얼굴 인식 기능을 배치하는 등 원격 식별 목적으로 생체인식 데이터를 수집하고 사용하는 것은 기본권에 특정한 위험을 수반한다. 원격 생체인식 AI 시스템 사용이 기본권에 미치는 함의는 사용 목적, 상황, 범위에 따라 상당히 달라질 수 있다.

EU 개인정보 보호법은 특정 조건을 제외하고는 자연인을 고유하게 식별하려는 목적으로 생체인식 정보를 처리하는 것을 원칙적으로 금지하고 있다. GDPR 하에서 구체적으로는, 그러한 처리가 몇 가지 근거에 의해서만 이루어질 수 있는데, 주된 경우는 상당한 공익상의 이유가 있을 경우이다. 이 경우, 처리는 비례성, 개인정보 보호권의 본질에 대한 존중 및

적절한 안전조치에 대한 요건에 따라 EU 법률 및 회원국 국내법에 기반하여 이루어져야 한다. 법집행기관 지침에 따르면, 이러한 처리에 대한 엄격한 필요성이 있어야 하며, 원칙적으로 적절한 안전조치 뿐만 아니라 EU 법률 및 회원국 국내법에 의한 허가도 있어야 한다. 자연인을 고유하게 식별할 목적으로 생체인식 정보를 처리하는 것은 EU 법률에 규정된 금지 조항에 대한 예외에 관한 사항이므로, EU 기본권 헌장의 적용을 받게 된다.

따라서, 현행 EU 개인정보 보호법 및 기본권 헌장에 따라, 그 사용이 절차적으로 정당하고, 비례적이며, 적절한 안전조치의 구비된 원격 생체인식 목적으로만 AI를 사용할 수 있다.

공공장소에서 그러한 목적을 위해 AI를 사용하는 것과 관련된 사회적 우려를 해소하고 내부 시장의 분열 방지하기 위해, 집행위는 그 사용을 정당화할 수 있는 특정 상황 및 공통적인 안전조치에 대해 유럽 내에서 광범위한 토론을 시작할 것이다.

E. 시정

위에서 언급한 고위험 AI 애플리케이션과 관련하여 적용될 법적 요건의 수범자와 관련하여, 고려해야 할 두 가지 주요 문제가 있다.

첫째로, 관련된 경제 행위자들 사이에서 의무가 어떻게 분배되어야 하는지에 대한 의문이 있다. AI 시스템의 생애주기에 많은 행위자들이 관여되어 있다. 여기에는 개발자, 배포자(AI가 탑재된 제품이나 서비스를 사용하는 사람) 및 잠재적으로 다른 이들(생산자, 유통업자 또는 수입업자, 서비스 제공자, 직업적 또는 개인적 사용자)이 포함된다.

집행위의 견해는 미래의 규제 프레임워크에서 잠재적 위험을 다루기에 가장 적합한 행위자에게 각 의무를 부여해야 한다는 것이다. 예를 들어, AI 개발자들은 개발 단계에서 발생하는 위험을 다루는데 가장 적합할 수 있지만, 사용 단계 동안 위험을 통제하는 능력은 보다 제한적일 수 있다. 이 경우, 배포자가 관련 의무를 부담해야 한다. 이런 접근법은 최종 사용자 또는 피해를 입은 기타 당사자들에 대해 책임을 지고 사법 수단에 대한 효과적인 접근을 보장하려는 목적에 따라, 야기된 손해에 대해 어떤

당사자가 책임을 져야 하는지에 대한 문제에서 선입견(prejudice)을 갖지 않는다. EU 제품 책임법에 따르면, 결함이 있는 제품에 대한 책임은 생산자에게 있는데, 이 경우 다른 당사자들에게 복구하도록 하는 국내법에 대한 악영향은 없다.

둘째, 입법 개입의 지리적 범위에 대한 의문이 있다. 집행위의 관점에서, 이 요건은 EU 역내에서 AI 탑재 제품이나 서비스를 제공하는 모든 관련 경제 행위자들에게 요구사항이 적용되는 것이 가장 중요하며, 이들이 EU에 설립되었는지 여부와는 관계가 없다. 그렇지 않으면 앞서 언급한 입법 개입의 목적이 완전하게 달성될 수 없었다.

F. 규정 준수 및 이행

AI가 신뢰할 수 있고, 안전하며, 유럽의 가치와 규율을 존중하도록 보장하기 위해, 적용 가능한 법적 요건들을 실제로 준수되어야 하며, 관할 국가 및 유럽 당국은 물론 관련 당사자들은 모두 이를 효과적으로 이행해야 한다. 관할 당국은 개별 사례를 조사할 수 있어야 하되 사회에 미치는 영향도 평가할 수 있는 위치에 있어야 한다.

특정 AI 애플리케이션이 시민과 우리 사회에 미치는 높은 위험(위의 섹션 A 참조)을 고려해 볼 때, 이 단계에서 집행위는 위험성이 높은 애플리케이션에 적용되는 상기의 의무적 요구사항들이 준수되는지 검증하고 보장하기 위해 객관적이고 사전적인 적합성 평가가 필요하다고 생각한다(섹션 D 참조). 사전 적합성 평가에는 테스트, 검사 또는 인증에 대한 절차가 포함될 수 있다. 여기에는 개발 단계에서 사용되는 알고리즘과 데이터셋에 대한 점검이 포함될 수 있다.

고위험 AI 애플리케이션에 대한 적합성 평가는 EU의 내부 시장에 배치되는 다수의 제품에 대해 이미 존재하는 적합성 평가 메커니즘의 일부여야 한다. 그러한 기존 메커니즘에 의존할 수 없는 경우 유사한 메커니즘을 수립할 필요가 있을 수 있는데, 이는 이해당사자들과 유럽 표준기구의 모범 관행과 가능한 참여에 기반하여 수립되어야 한다. 그러한 새로운 메커니즘은 비례적이고 차별적이지 않아야 하며 국제적 의무를 준수하는 투명하고 객관적인 기준을 사용해야 한다.

사전 적합성 평가에 의존하는 시스템을 설계하고 구현할 때, 다음과 같은 사항에 대해 특별히 고려해야 한다.

- 위에서 설명한 모든 요구사항이 사전 적합성 평가를 통해 검증되는 것은 적절하지 않을 수 있다. 예를 들어, 제공되어야 할 정보에 대한 요구사항은 일반적으로 그러한 평가를 통한 검증에 적합하지 않다.
- 특정 AI 시스템이 진화하고 경험에서 학습할 가능성을 특히 고려해야 하는데, 이 경우 해당 AI 시스템의 수명에 대해 반복적인 평가가 필요할 수 있다.
- 훈련에 사용된 데이터는 물론, AI 시스템을 구축, 테스트 및 검증하는 데 사용된 관련 프로그래밍 및 훈련 방법론, 절차 및 기법을 검증할 필요가 있다.
- 적합성 평가 결과 AI 시스템이 요구사항을 충족하지 못하는 것으로 나타나면 확인된 단점을 보완해야 한다. 예를 들어 훈련하는 데 사용된 데이터 등과 관련한 요구사항이 충족되지 않으면 모든 해당 요구사항을 충족하는 방식으로 EU 내에서 시스템을 다시 훈련하여 보완해야 한다.

적합성 평가는 요구사항에서 다루는 모든 경제 행위자들에게 설립 지역과 관계없이 의무적일 것이다. 중소기업의 부담을 제한하기 위해 디지털 혁신 거점을 비롯해 지원 구조를 어느 정도 구상할 수 있다. 또한, 표준은 물론 온라인 전용 도구가 규정 준수를 촉진할 수 있다.

모든 사전 적합성 평가는 관할 국가 기관이 준수 및 사후 시행 여부를 모니터링할 때 선입견이 없어야 한다. 이는 고위험 AI 애플리케이션과 관련하여 사실일 뿐 아니라, 법적 요구사항이 적용되는 다른 AI 애플리케이션에 대해서도 해당된다. 비록 문제의 애플리케이션이 가진 고위험 속성은 관할 국가 기관이 전자에 특별히 주의를 기울여야 하는 이유가 될 수 있겠지만 말이다. 사후 통제는 관련 AI 애플리케이션에 대한 적절한 문서화에 의해 가능하며(위의 섹션 E 참조) 적절한 경우 관할 당국과 같은 제3자가 그러한 애플리케이션을 테스트할 수도 있어야 한다. 이는 기본권에 위험이 발생하는 경우에 특히 중요할 수 있는데 상황에 따라 다르다. 이러한 준수 모니터링은 지속적인 시장 감시 계획의 부분이 되어야 한다. 거버넌스 관련 측면은 아래 섹션 H에서 추가적으로 논의한다.

또한 고위험 AI 애플리케이션과 기타 AI 애플리케이션 모두에서, AI 시스템으로부터 부정적 영향을 받은 당사자에 대한 효과적인 사법적 보상이 보장되어야 한다. 책임과 관련된 이슈는 본 백서에 첨부된 안전 및 책임 프레임워크에 대한 보고서에서 추가로 논의된다.

G. 무위험 AI 애플리케이션을 위한 자율 표시 제도

'고위험'으로 분류되지 않아(위의 섹션 C 참조) 위에서 논의한 의무적 요구사항에 해당하지 않는 AI 애플리케이션의 경우(위의 섹션 D, E, F 참조), 법률 적용 외의 선택지로서 자율 표시 제도 수립을 들 수 있다.

이 제도 하에서, 이해관계가 있지만 의무 요구사항에 포함되지 않는 경제 행위자들은 그러한 요구사항들이나 자율적 차원에서 수립된 특정한 유사 요구사항에 대해 자율적으로 적용하기로 결정할 수 있다. 그 후 관련 경제 행위자들은 자사 AI 애플리케이션에 대해 품질 표시를 수여받게 된다.

자율 표시 제도는 관련 경제 행위자들에게 있어 자사 AI 탑재 제품과 서비스가 신뢰할 수 있다는 신호가 될 수 있을 것이다. 이는 사용자들로 하여금 문제의 제품과 서비스가 특정 목적 및 표준화된 EU 전체 기준을 준수하고 있음을 쉽게 인식할 수 있게 할 것이고, 이는 통상적으로 적용되는 법적 의무를 넘어서는 것이다. 이를 통해 AI 시스템에 대한 사용자의 신뢰도를 높이고 전반적인 기술 활용을 촉진할 수 있다.

이 선택지는 고위험으로 간주되지 않는 AI 시스템의 개발자 및 배포자에 대한 자율 표시 체제 수립이라는 새로운 법적 도구의 창출을 수반한다. 표시 제도에 참여하는 것은 자율적이지만, 개발자나 배포자가 표시를 사용하기로 선택한 이상 요구사항이 구속력을 갖게 된다. 시행 전과 시행 후의 조합은 모든 요구사항에 준수를 보장해야 할 것이다.

H. 거버넌스

책임구조의 분열을 방지하고 회원국의 역량을 증대시키며 유럽이 AI 탑재 제품 및 서비스 테스트 및 인증에 필요한 역량을 점진적으로 갖추기

위하여, 각국 관할 기관 간의 협력 프레임워크 형태의 유럽 AI 거버넌스 구조가 필요하다. 이런 차원에서, AI가 사용되는 곳에서 관할 국가 기관들이 자신의 권한을 발휘할 수 있도록 지원하는 것이 유익할 것이다.

유럽 거버넌스 구조는 정보 및 모범 관행을 정기적으로 교류하고, 신흥 동향을 파악하며, 인증뿐 아니라 표준화 업무에 대해 조언하는 포럼으로서 다양한 직무를 수행할 수 있다. 또한 지침, 의견, 전문자료 발행을 통해 법률 체제의 이행을 촉진하는 데 핵심적인 역할을 해야 한다. 그러한 효과를 위해, 이는 회원국 및 EU 차원의 부문별 네트워크와 규제 당국뿐만 아니라 국가 당국간 네트워크를 필요로 한다. 게다가, 전문가들로 구성된 위원회는 집행위에 도움을 제공할 수 있다.

거버넌스 구조는 이해당사자 참여를 최대한 보장해야 한다. 소비자 단체 및 사회적 대화자(social partner, 노・사), 기업, 연구자 및 시민단체 등 이해당사자들은 프레임워크의 구현과 추가 개발에 대해 상의할 필요가 있다.

금융, 제약, 항공, 의료기기, 소비자 보호, 개인정보 보호 등 기존 구조를 고려해보았을 때 이번에 제안되는 거버넌스 구조가 기존 기능과 중첩되어서는 안 된다. 그 대신에 기존의 전문지식을 보완하고 기존 당국이 AI 시스템과 AI 탑재 제품 및 서비스에 관여한 경제 행위자들의 활동을 감시하고 감독할 수 있도록 지원하기 위해 다분야 여러 EU 및 국내 관할 당국과 긴밀한 연계를 구축해야 한다.

마지막으로, 이 선택사항이 추진될 경우, 적합성 평가의 수행은 회원국이 지정한 인증 기관에 위탁할 수 있다. 테스트 센터는 위에서 설명한 요구사항에 따라 AI 시스템에 대한 독립적인 감사 및 평가가 가능해야 한다. 독립적인 평가는 신뢰를 높이고 객관성을 보장할 것이다. 그것은 또한 관련 관할 당국의 업무를 용이하게 할 수 있다.

EU는 우수한 테스트 및 평가 센터들을 보유해 왔으며 AI 분야에서도 역량을 키워야 한다. 내부 시장 진입을 원하는 제3국 설립 경제 행위자들은 EU에 설립된 인증 기관을 이용하거나 제3국과의 상호인정협정에 따라 그러한 평가를 수행하도록 지정된 제3국 기구를 이용할 수 있다.

AI와 관련된 거버넌스 구조와 여기서 문제되는 해당 적합성 평가는 현행 EU 법률에 따라 특별 분야 관할 당국이 특정 문제(금융, 제약, 항공,

의료기기, 소비자 보호, 개인정보 보호 등)에 대해 가지고 있는 권한과 책임에 영향을 주지 않을 것이다.

6. 결론

인간중심적이고 윤리적이며 지속가능하며 기본적 권리와 가치를 존중한다면 AI는 시민과 기업, 사회 전반에 많은 혜택을 가져오는 전략 기술이다. AI는 유럽 산업의 경쟁력을 강화하고 시민의 복지를 향상시킬 수 있는 효율성과 생산성 측면에서 중대한 이점을 제공한다. 그것은 또한 기후변화와 환경파괴와의 싸움, 지속가능성과 인구변화와 관련된 문제, 그리고 민주주의 수호, 그리고 필수적이고 비례적인 한도에서 범죄와의 싸움 등 가장 시급한 사회문제들에 대한 해결책을 찾는 데 기여할 수 있다.

유럽이 AI가 제공하는 기회를 완전히 잡으려면 필요한 산업·기술 역량을 개발하고 강화해야 한다. 유럽 데이터 전략에서 제시된 바와 같이 EU가 글로벌 데이터 허브가 될 수 있도록 지원하는 조치도 필요하다.

유럽식 AI 접근법은 EU 경제 전반에 걸쳐 윤리적이고 신뢰할 수 있는 AI의 개발과 활용을 지원하는 동시에 AI 분야에서 유럽의 혁신역량을 촉진하는 것을 목표로 하고 있다. AI는 사람들을 위해 일하고 사회 공동선에 힘이 되어야 한다.

본 백서 및 부속 안전 및 책임 프레임워크에 대한 보고서와 함께, 유럽 집행위원회는 AI에 대한 유럽식 접근법에 대한 구체적인 제안에 대한 회원국 시민사회, 산업계 및 학계의 광범위한 의견 청취를 시작한다. 여기에는 연구와 혁신에 대한 투자를 활성화하기 위한 정책 수단이나 중소기업의 AI 기술개발 및 활용을 지원하기 위한 정책 수단, 미래 규제 프레임워크의 핵심 요소에 대한 제안 등이 모두 포함된다. 이러한 협의 과정은 관련 당사자들과의 포괄적 의견교환으로 집행위의 다음 단계에 유용할 것이다. □

영국 AI 조달지침[61)]

Guidelines for AI procurement

●

2020. 6.
영국 비즈니스・에너지・산업전략부,
디지털・문화・미디어・스포츠부, 인공지능실

도입

AI란 무엇이며 정부는 AI를 어떻게 활용할 수 있을까

인공지능(AI)은 비용을 절감하고 품질을 향상시키며 일선 직원들의 소중한 시간을 절약함으로써 공공 서비스를 크게 개선할 수 있는 일련의 기술들로 구성되어 있다.

> AI에 대해 <공공부문 AI 활용 가이드>에서는 다음과 같이 정의하고 상세하게 설명하였다.
>
> "AI는 일반적으로 지능이 필요하다고 생각되는 작업을 수행할 수 있는 시스템을 구축하기 위한 디지털 기술의 이용으로 정의할 수 있다."
>
> AI의 발전은 끊임없이 진화하고 있지만, 일반적으로 통계를 사용하여 대량의 데이터에서 패턴을 찾아내고 그 데이터를 사용하여 인간의 지속적인 지도 없이 반복적인 작업을 수행하는 기계들을 의미한다.

61) 2020년 6월 영국 비즈니스・에너지・산업전략부, 디지털・문화・미디어・스포츠부, 인공지능실은 공동으로 <AI 조달 지침>을 발표함. <https://www.gov.uk/government/publications/guidelines-for-ai-procurement/guidelines-for-ai-procurement>.

우리는 정부에 AI 시스템을 배치하는 초기 단계에 있다. 우리는 AI 시스템을 사용하여 의사결정을 진행할 때의 새로운 편익을 계속 발견하고 있으며, 동시에 우리가 다루어야 할 문제와 위험 또한 계속 발견되고 있다.

이 지침은 대부분 머신러닝(machine learning)의 사용을 염두에 두고 있다. 머신러닝은 AI의 하위집합이며, 주어진 과제에 대해 시간을 경과하며 경험을 통해 성능을 향상시키는 디지털 시스템을 개발하는 것을 말한다. 머신러닝은 AI의 가장 널리 사용되는 형태로서 자율주행차, 음성인식, 기계 번역과 같은 혁신에 기여해 왔다.

> 더 알고 싶다면
>
> AI 분야에서는 새로운 개념들이 많이 사용되고 있으며 별첨의 AI 용어집을 참고하는 것이 유용할 수 있다. 영국 공공부문에 AI가 어떻게 활용돼 왔는지에 대한 자세한 사례는 <공공부문 AI 활용 가이드>의 사례연구를 살펴볼 수 있다. AI 기술에 대해 보다 자세히 알아보려면, 국방과학기술연구소 AI랩의 데이터과학 전문가들이 국방부 이용자들을 위해 AI, 데이터과학, 머신러닝 등을 이해할 수 있도록 편찬한 비스킷북이 나와 있다.

이 지침의 목적은 무엇인가?

공공조달은 AI의 채택을 촉진할 수 있고 공공서비스 전달을 개선하는 데 기여할 수 있다. 정부의 구매력은 이러한 혁신을 주도하고 영국의 AI 기술 개발을 촉진할 수 있다.

AI가 신흥 기술인 만큼 공공기관의 요구사항에 맞는 최적의 시장 접근 경로를 구축하는 것이 어려울 수 있으며, 혁신적인 공급업체와 실제 거래하거나 AI 기술을 효과적이고 윤리적으로 알맞게 배치하도록 AI 특수 기준과 약관을 마련하는 것이 어려울 수도 있다.

지침은 AI 기술을 구매하는 방법에 대한 이행 원칙들과 더불어 조달 과정에서 발생할 수 있는 과제를 해결하는 통찰력을 제공할 것이다. 이 지침은 이러한 지침들 중 첫 번째이며, 총망라한 것은 아니다.

지침은 인공지능실이 세계경제포럼 4차산업혁명센터, 정부 디지털서비스청, 내각 사무처 정부조달부서(Government Commercial Function), 공공조달청(Crown Commercial Service) 등과 공동으로 개발한 것이다. 산업계, 학계 및 정부 부처의 광범위한 이해관계자들이 지침 개선에 도움을 주었다. 이 지침은 세계경제포럼의 '공공부문 AI 잠금 해제' 프로젝트에서 출발했다.

또한 인공지능실은 이 프로젝트의 일환으로 <AI 조달키트>를 공동 제작했다. 이는 전세계 공공부문 조달 전문가들이 AI 조달에 대한 접근방식을 재고할 수 있도록 지원하는 툴킷이다.

> 이 지침은 정부의 AI 기술 활용이 진화할 때마다 업데이트될 예정이다. AI 시스템을 조달 중이거나 조달하는 것을 고려하고 있는 경우, 혹은 지침에 대한 피드백을 제공하는 데 관심이 있는 경우 ai-properties-guidelines@officeforai.gov.uk으로 문의 바란다.

누가 이 지침을 사용해야 하는가?

이번 가이드라인은 기존 서비스를 개선하려는 목적으로나 향후 서비스 전환의 일환으로 AI 기술의 적합성을 검토 중인 중앙부처들을 대상으로 한다. 다른 공공부문 기관들도 이 지침을 따를 수 있다.

이 지침은 다음과 같은 이들에게 유용할 것이다.

- AI 기술 조달 및 계약 관리를 담당하는 조달 및 상업 부문 실무자
- 기술적 문제를 해결하기 위해 AI 시스템의 적합성을 검토 중인 데이터, 정보, 기술 및 혁신 부문 최고책임자
- 디지털 변환 프로젝트 및 프로그램에 AI 기술을 사용하고자 하는 디지털 전달 및 변환팀
- AI 시스템의 프로젝트 특수 요구사항을 준비 중이거나 이를 평가, 사용, 관리하는 분석가, 데이터 과학자 및 기타 디지털・데이터・기술 전문가

- 정부 내 AI 조달에 대한 모범 관행, 기술 및 윤리적 표준을 보다 잘 이해하고자 하는 공급업체

공공조달은 조달 규칙 및 규정 프레임워크에 의해 관리되며, 이 지침은 독자가 이러한 규칙과 단대단 조달 절차에 대한 충분한 실무지식을 가지고 있다고 가정한다. 경영상 판단 속에 본 지침을 사용하고, 적절한 경우 법률 자문을 구하기 바란다. 일부 연구용역 계약은 조달 법령의 범위를 벗어날 수도 있다.

이 지침을 어떻게 사용해야 하는가?

이 지침은 AI 시스템의 실행가능성을 평가할 때 고려해야 할 주제를 비롯해, AI 기술 프로젝트를 집행할 때 조달팀이 고려해야 할 사항도 개괄적으로 제시한다. 이행 원칙으로서, 가능한 한 개방적으로 작업하면서 AI 프로젝트와 사용하게 될 툴, 데이터 및 알고리즘에 대해 투명할 것을 요구한다.

모든 형태의 AI 시스템이 동일하지는 않을 것이며 AI 기술이 점점 더 많은 종류의 기술 제품에 내장되는 형태일 것이다. AI 시스템은 처음부터 개발하거나, 매대에서 구매하거나, 이미 사용 중인 시스템에 추가할 수도 있다.

공급업체가 AI가 아닌 특수 요구사항 또는 시스템 제공의 일부로 AI 기술을 활용할 것을 제안하는 경우, 몇 가지 추가적으로 고려해야 할 사항이 있을 수 있다. 이러한 경우, 디지털 정보 최고 책임자/기술 설계자와 상의하여 AI 모델이 솔루션에 미칠 수 있는 영향을 평가하고, 계약에 AI 기술을 제공의 일부로 수용하는 적절한 조항을 규정하기 위해 경영상 판단을 내려야 한다.

이 지침은 기술 및 디지털 서비스의 사용과 관련하여 다음과 같은 기존 정책 및 지침과 함께 고려되어야 한다.

- 디지털 서비스 표준
- 기술 규칙 (The Technology Code of Practice)

- 데이터 윤리 프레임워크
- 공공부문 AI 활용 가이드
- 데이터 개방 표준
- 기타 기술 표준 및 지침

<아웃소싱 플레이북>을 참조하여, AI 구매 전략을 정의할 때 다른 기술 요구사항에 대한 구매 전략과 동일한 방식으로 한다.

> 더 알고 싶다면
>
> <공공부문 인공지능 활용 가이드>는 정부 팀으로 하여금 AI 시스템을 활용해 문제를 어떻게 해결할 수 있는지 이해하고, 개발이 필요한 해결책의 일부로 AI 기술이 사용될 수 있을지 의사결정을 내릴 때 도움을 줄 수 있다. <AI 조달키트> 툴킷은 AI 조달에 대한 상세한 지침은 물론, 절차 중 고려해야 할 주요 이슈에 대해 보다 상세한 내용을 제공한다. 또한 NHSx에서 발행한 <보건의료 구매자 AI 체크리스트>와 같은 부문별 지침을 참고할 수 있다.

10가지 우선 고려 사항

1. AI 도입 계획에 조달을 포함할 것

AI 기술 도입을 접목하도록 기술 및 데이터 계획을 갱신할 것. 정부 전반적으로 AI 도입을 지원하는 조달 정책을 전략적으로 사용하고, 협업을 통하여 AI 기술 배치에 있어 규모의 경제를 활용하는 이점을 취하도록 하며, 정부 전반에 걸쳐 관심 있는 팀들과 지식을 공유할 것.

AI 관련 이니셔티브를 주도하는 중앙 정부 부처 및 기관내 다른 팀들과 업무를 연계하는 방안을 고려해 볼 것.

기관 내부 및 민간 서비스 전반에 걸쳐 네트워크를 구축하여 통찰력을 공유하고 모범 관행을 통해 학습할 것.

2. 다양한 다학제 팀에서 의사를 결정할 것

AI 프로젝트의 개발, 평가, 전달 시 AI 기술이 접목된 상호의존적 분야를 이해하는 다양한 팀과 함께 하면 더욱 효과적임. 여기에는 다음이 포함될 수 있음.

- 분야별 전문지식(예: 의료, 운송)
- 상업적 전문지식
- 시스템 및 데이터 엔지니어링
- 모델 개발(예: 심층 학습)
- 데이터 윤리학
- 시각화/정보 설계

낙찰된 공급업체에게, 적합한 기술력을 갖춘 팀을 구성하고 AI 시스템의 편향성을 완화하기 위해 다양성에 대한 요구를 해결할 것을 요구할 것.

3. 조달 절차 개시 전 데이터 평가를 실시할 것

데이터는 현재 AI 기반 솔루션 대다수의 기반임. 관련 데이터의 가용성은 모든 AI 시스템의 전제 조건이기 때문에 가용 데이터가 없다면 AI 조달에 대해 논의하는 데 시간을 낭비할 이유가 없음.

- 조달 절차의 개시 단계부터 데이터 거버넌스 메커니즘이 가동될 수 있도록 확보할 것.
- 프로젝트에 관련 데이터를 사용할 수 있는지 여부를 평가할 것.
- 시장에 출시하기 전에 데이터 내부의 결함 및 편향 가능성을 해결할 것. 데이터 문제를 직접 해결할 수 없는 경우 이를 해결하기 위한 계획을 수립할 것.
- 조달 계획 및 후속 프로젝트를 위해 공급업체와 데이터를 공유할 것인지 여부 및 방법을 정의할 것.

4. AI 배치의 장점 및 위험성을 평가할 것

공익적 목표를 정의하는 것은 AI 시스템이 달성하고자 하는 프로젝트 및 조달 절차 전반의 기반이 됨. 또한 AI 기술은 조달 단계 초기에 확인하고 관리해야 하는 특정한 위험성을 야기함.

- 제안서를 평가할 때 공익이 의사결정 절차의 주요 동인이라는 점을 조달 문서에 설명할 것. <사회적 가치> 지침에 따라 AI 시스템이 인간과 사회 경제에 미치는 영향 및 편익을 고려할 것. 조달되고 있는 사업이 (본질적으로 일반적이지는 않다 하더라도) 공익적 목표와 관련이 있어야 하며, 차별금지, 동등한 대우 및 비례성의 원칙을 준수해야 함.
- 당면한 문제와 관련하여 AI를 고려한 배경을 조달 문서에 명확히 설명하고 대안적 솔루션에 대해 열린 태도를 취할 것.
- 조달 절차 개시 단계에서 AI 영향 평가를 수행하고, 중간 조사 결과가 조달에 반영되는지 확인할 것. 주요 의사 결정 단계에서 평가 결과를 재차 살펴볼 것.

5. 처음부터 사실상 시장에 참여할 것

정부 지출은 공정하고 경쟁적인 시장을 만드는데 사용될 수 있고, 이는 더 나은 AI 시스템으로 이어짐. AI 공급업체와의 조기 협력은 보다 관련성이 높은 대응으로 이어져 성공적인 조달 및 보다 나은 프로젝트 수행 가능성을 높일 수 있음. 접근 방식의 비례성에 초점을 맞추어 스타트업, 중소기업, 자율・지역・사회적 기업 등의 공급업체는 물론 경쟁에서 대표성이 낮은 그룹이 소유한 공급업체 등을 단념시키는 불필요한 부담을 부과하지 말 것.

- 계획 단계 초기 및 과정 중에 AI 공급업체와 협력할 것.
- 다양한 방법으로 광범위한 AI 공급업체를 접촉할 것.
- AI 생태계의 경쟁을 지원하는 개방적 환경을 장려할 것.

6. 알맞은 시장 접근 경로를 수립하고, 특정 솔루션보다는 과제에 초점을 맞출 것

조달되는 AI 시스템은 문제가 되고 있는 과제를 해결하고 책임 있고 혁신적인 시장의 반응을 촉진해야 함. 신중하게 작성된 요구사항은 공급업체로 하여금 귀 기관이 필요로 하는 사항을 이해하고 최선의 솔루션을 제안하는 데 도움이 될 수 있음. 공급업체에게 상황이나 과제에 대해 알려서, 공급업체들이 귀 기관의 요구에 맞는 솔루션을 제안하도록 할 것.

- 상업적 모범 관행은 <아웃소싱 플레이북>의 지침을 참조할 것.
- <혁신 파트너십>, GovTech Catalyst, 공공조달청의 동적 AI 구매 시스템 등 AI 시스템을 구할 수 있는 다양한 시장 경로를 탐색해 볼 것.
- 솔루션에 대한 자세한 사양 대신 명확하게 문제를 진술할 것.
- 제품 개발에서 반복적인 접근방식의 경우 우선순위를 정하고 이를 입찰 공고에 반영할 것.

7. 거버넌스 및 정보 인증을 위한 계획을 수립할 것

AI 시스템의 생애주기 전체에 걸쳐 정밀 조사가 가능하도록 적절한 감독 메커니즘을 수립해야 함. AI 사용 사례와 프로젝트의 위험 요인에 따라 다른 고려사항을 적용하고, 이러한 접근이 정밀조사를 감당할 수 있는지 확인해야 함. 조달 문서에서 기존 법과 규정을 준수하고 규범의 표준화를 지원해야 한다는 점을 강조할 것.

요구사항 초안을 작성할 때 기존 규칙, 지침 및 규정을 반드시 참조하고, 해당되는 경우 이들을 계약 조건에 반영할 것.

- <기술 규칙> 및 <정부 설계 원칙>, <데이터 윤리 프레임워크> 및 기타 관련 표준을 준수할 것.
- AI 의사 결정의 투명성을 최대화하여 사용자에게 AI 시스템이 잘 기능한다는 확신을 부여할 것.

8. 블랙박스 알고리즘 및 공급업체 종속(Lock-In)을 방지할 것

알고리즘의 설명 가능성과 해석 가능성을 장려하고 이를 설계 기준 중 하나로 삼을 것. 이는 귀 기관 팀이 그 결과를 이해하는 것을 가능케 하는 방법 및 기술이 사용된다는 것을 의미함. 고도로 '설명 가능한' AI 시스템의 산출물은 귀 기관 팀은 물론 다른 공급업체에 의해 해석될 수 있음. 이는 또한 귀 기관이 향후 AI 시스템을 지속하거나 구축할 때 다른 공급업체와 협력할 수 있도록 함으로써 공급업체에 종속될 수 있는 위험을 제한함.

9. 평가 시 AI 배치의 기술적, 윤리적 한계를 해소해야 할 필요성에 초점을 맞출 것

다학제 팀의 경험을 활용하여 평가 절차를 지원하고 입찰 평가를 수행할 때 광범위한 전문 지식을 확보할 것.

- 공급업체가 데이터 내에서 편향성 문제에 주목하거나 해결하였는가? 왜 그들의 전략이 적절하고 비례적인지 명확하게 설명하였는가? 귀 기관이 놓쳤을 수 있는 문제에 대해 공급업체가 이를 해결할 계획을 가지고 기민하게 프로젝트를 제공하는 문제의 중요성을 강조했는가?
- 기존 서비스 또는 기술과의 통합 필요성을 고려하였는가?
- 공급업체의 거버넌스 접근 방식이 귀 기관의 요구사항을 충족하는가?
- 적절한 기술 표준을 준수하였는가?

10. AI 시스템의 생애주기 관리를 고려할 것

공공부문의 AI 기반 솔루션이 윤리적인 사용을 보장하기 위해서는 집행 계획, 지속가능하고 계속적인 평가 방법, 데이터 모델에 대해 피드백하는 메커니즘이 중요함. 더불어 AI 시스템의 기능성 및 결과는 조달 절차에서 뚜렷하지 않을 수 있으며, 배치 과정에서야 드러나는 경우가 많아 구매 기관과 공급업체 간 소통 및 정보 공유가 확대되어야 함.

- AI 조달 과정에서 일회성 결정이 아니라 생애주기에 걸친 테스트가 필요하다는 점을 고려할 것.
- 지식 이전 및 교육훈련을 요구사항의 일부로 포함할 것.
- AI 시스템을 이해해야 하는 비전문가를 대상으로 교육훈련 및 설명을 실시할 것을 요구사항의 일부로 포함할 것.
- 적절하고 지속적인 고객지원 및 호스팅 협의가 이루어지도록 보장할 것.

조달 절차 내 AI 특수 조건 고려사항

이 장에서는 조달 절차 전반에 걸쳐 다루어야 할 구체적인 고려사항을 제기한다.

1. 준비 및 계획
2. 공고
3. 선정, 평가 및 낙찰
4. 계약 이행 및 지속적인 관리

일반적인 원칙으로서, 모든 AI 조달에 대한 조사는 “어떻게 하면 우리의 문제를 AI 시스템 솔루션에 맞출 수 있을까?”가 아니라 “AI 기술이 우리에게 어떤 혜택을 줄 수 있을까?”라는 발상으로 이루어져야 한다.

AI 기술을 다른 기술 솔루션과 마찬가지로 취급하고 적절한 상황에서 사용하라. 모든 서비스는 서로 다르게 마련이며, 기술에 대해 내리는 귀 기관의 결정은 서비스별로 특수할 것이다.

1. 준비 및 계획

모두에게 개방적이고 접근 가능하여 광범위한 참여를 장려하는 유연하고 효율적인 조달 절차를 달성하기 위해서는 준비가 핵심적이다. 가능한 한 개방적으로 일하며, 사전 조달의 법적 요건을 준수해야 한다. 여기에는 2012년 「공공서비스(사회적 가치)법」(해당되는 경우 개정법)에서 요구하는 의무적 고려사항 반영과 2010년 「평등법」에 따른 <공공부문 평등 의무>의 적용성 평가 등이 포함된다. AI 시스템 배치를 포함할 가능성이 있는 조달 프로젝트를 시작하기 전에 다음 사항을 고려해야 한다.

다학제 팀:

AI가 실행 가능하고 적절한 솔루션인지 여부를 검토하기 위해 지식과 경험을 가진 사람들로 팀을 구성할 것. AI 시스템의 조달 및 구현을 지원하기 위해 다양한 역할과 기술이 결합된 다학제 팀의 구성을 추구할 것. 다양한 기술을 갖춘 팀은 귀 기관이 데이터 및 영향 평가를 수행하는 것에 기여하고 귀 기관의 사업 사례 및 조달 절차에 주요 발견 사항들을 반영하는 데 기여할 수 있음.

AI 프로젝트 팀에서 고려해야 할 전문적 역할들:

- 데이터 설계자
- 데이터 과학자
- 데이터 엔지니어
- 기술 설계자
- 제공 관리자
- 보안 설계자
- 상업화 관리자

처음부터 이 모든 역할이 필요하지 않을 수도 있지만, 시작하기 전에 자신의 필요를 고려할 필요가 있음. 시장에 진출할 수 있는 적절한 기반이 마련되어 있는지 검토하고, AI 시스템을 기존 절차, 기술, 서비스에 통합할 수 있는지 검증하기 위해서는 전문가와 상담하는 것이 유용함.

견고한 실행이 이루어지고, 팀의 기술력 내에서 작업이 수행되도록 하는 것이 중요함. 팀에 전문 지식이 부족한 경우 기관 또는 정부 내 전문가 네트워크를 접촉하여 원하는 사용 사례에 대한 중요한 통찰력을 얻을 수 있음.

또한 공급업체의 요구사항 충족을 검토하는 의사결정에서 인재 발굴 활동의 완수 여부를 고려할 수 있음.

> 더 알고 싶다면
>
> 컨설팅에 유용할 수 있는 팀과 기관의 예시로는 인공지능실, 정부 디지털 서비스, 데이터 윤리 및 혁신센터 또는 특정 부문 지식을 갖춘 팀과 기관들이 있다. <공공부문 AI 활용 가이드>에서도 AI 활용 사례를 찾아볼 수 있다. 또한 지식 허브, 디지털 구매 커뮤니티, 데이터 과학 이해관계자 커뮤니티 또는 기타 유사 네트워크를 비롯한 전문가 커뮤니티를 통해 모범 관행을 알아보고 지식과 피드백을 공유할 수 있다.

데이터 평가 및 거버넌스

시장에 진출하기 전에 귀 기관의 데이터를 찾아서 확보해둘 것. 이를 위해서는 특수하고 전문적인 조언을 받아야 할 수도 있음. 데이터 설계자 및 데이터 과학자가 이 과정을 이끌어야 함. 이러한 과정은 귀 기관과 귀 기관의 팀이 귀 기관이 이용할 수 있는 데이터의 복잡성, 완전성 및 한계를 이해하는 데 도움을 줄 것임.

데이터에 대한 철저한 평가가 어려운 것으로 드러나거나 이루어지지 않은 경우, AI 시스템이 자기 의사결정의 기반으로 사용할 데이터에 대해 종합적인 점검을 실시할 것을 입찰공고 요구사항에 포함할 것.

데이터 거버넌스는 프로젝트와 관련된 모든 데이터 활동을 포함해야 함.

- 프로젝트 구성원에게 데이터 접근 권한 부여
- 분석을 위해 데이터를 다른 위치에 저장
- 데이터 [처리에 대한] 동의 검토

프로젝트와 데이터의 민감도에 따라 조달 절차에서 공급업체들에게 데이터를 공개하는 것도 고려해 볼 만함. 공급업체는 이를 통해 이용 가능한 데이터에 대해 파악하고 입찰 공고에 대한 대응을 개선할 수 있음. 모든 공급업체에 동일한 데이터를 동시에 제공하고 이때 개인정보 보호법 및 GDPR을 준수하는 조치를 취하고 있는지 확인할 것. 이를 지원하기 위해 NDA(이면계약) 또는 공급업체 참여 행사를 고려해볼 것. <데이터 윤리 프레임워크> 원칙 3은 데이터 거버넌스 및 비례성에 대한 추가 정보를 규정하고 있음.

또한 데이터 총계처리, 마스킹 및 합성을 포함해 개인정보를 보호하는 익명화 기법의 사용을 고려해 볼 수 있음. 입찰 공고는 데이터를 덜 침해적으로 사용하거나 덜 민감한 데이터셋을 이용하여 동일하거나 유사한 결과를 달성하는 혁신적인 기술 접근법을 장려해야 함.

AI 영향 평가

AI 영향 평가는 프로젝트 설계 단계에서 시작되어야 함. 솔루션 설계 및 조달 절차는 평가에서 확인된 위험성의 완화를 추구해야 함. 취득하게 될 AI 시스템의 사양을 모르는 상태에서 완전한 평가 실시가 불가능하기 때문에 AI 영향평가는 반복적인 과정이 되어야 함.

AI 영향 평가는 다음을 평가함

- 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

관련 위험성과 각각의 완화 전략이 영향 평가 내에서 규정되고 합의되

어야 하며, 이 전략은 해당되는 경우 '계속진행/중단' 등 주요 의사결정 시점을 포함해야 함. 이러한 결정 시점 또는 AI 시스템 설계에 상당한 변화가 있을 때마다 영향 평가를 검토할 것.

> 더 알고 싶다면
>
> 개인정보 보호 영향 평가 및 평등 영향 평가는 의도하지 않은 결과 가능성을 평가하기 위한 유용한 출발점을 제공할 수 있다. 자동화된 의사결정을 위한 위험 평가 문항의 예시로는 캐나다 정부의 <자동화된 의사결정에 대한 지침>과 AI Now의 <알고리즘 영향 평가 프레임워크>를 참조.

예비적 시장 협력

예비적 시장 협력은 AI가 해결책의 일부가 될수 있을지 여부 및 그 방법을 이해하는 데 도움이 됨. 예비적 시장 협력으로 알게 된 내용을 통해 문제 진술을 더 잘 정의할 수 있으며, 요구사항의 범위와 실현 가능성을 결정하는 데도 도움이 될 수 있음.

예비적 시장 협력은 전국적으로 AI 설계 및 전달에 전문적인 중소기업과 사회적 기업을 포함하여 서비스 전달 개선에 도움이 될 수 있는 공급업체를 적극적으로 발굴해야 함.

모든 예비적 시장 협력은 공공 조달의 원칙을 준수해야 하며 공급업체가 우선적 이익을 얻지 못하도록 처리되어야 함. 실무적으로 이는 특정 솔루션 또는 공급업체에 맞춘 기술 사양을 설정하지 않고, 이때 공유된 정보는 조달 절차 중에도 이용할 수 있도록 함을 의미함.

> 더 알고 싶다면
>
> 예비적 시장 협력 또는 발굴 단계(discovery phase)를 완료하면 서비스 구매나 구축 전에 문제를 이해하는 데 도움이 될 것이다. AI 실시 계획 및 준비에 대한 <공공 부문 AI 활용 가이드>에서 발굴 단계 조직화 방법에 대한 내용 참조.

조달 접근 방식 및 수단

AI 시스템 구매를 위해 시장에 접근하는 경로는 현재 다수 존재함. 어떤 종류의 과제를 해결해야 하는지에 따라 달라짐. G-Cloud, Digital Outcomes and Specialists, the Spark Dynamic Purchasing System 등 프레임워크 협약들이 고려해봄직한 유용한 출발점이 될수 있음.

혁신 중심의 조달 절차는 정부 내 신기술 채택을 가속화하고, AI의 혁신과 윤리적 발전을 촉진할 수 있는 기회를 제공함. 여기에는 다음이 포함될 수 있음.

- 다양한 단계에서 시장에 진출할 수 있고 솔루션이 실행되기 전에 기술을 테스트하는 개념 증명(proof-of-concepts)을 포함하는 신속 조달 절차. 발굴 단계 또는 개념 증명을 통해 AI 시스템이 광범위한 요구사항을 충족할 수 있는지 여부를 시연해볼수 있음.
- 기술 경연대회, 시연회, 과제 기반 조달 절차 등은 공급업체들이 AI 기술을 바탕으로 서로 경쟁하게 하고, 기관이 해결되길 바라는 과제에 적용되는 기술을 평가해볼수 있음. 이러한 절차들은 혁신에 초점을 맞추고 다양한 접근 방식을 탐구해볼 수 있음. 예를 들어 GoverTech Catalyst 또는 스코틀랜드 정부가 운영하는 CivilTech® 액셀러레이터 프로그램을 들 수 있음.
- 혁신 파트너십은 현재 시장에서 접근가능한 옵션으로는 제공할 수 없는 기술의 조달을 가능케 함. 공공 계약 개정 규정은 이 시장 접근 경로가 잠긴 문이 열 때 나타나는 기회들을 강조함. 한 연구보고서는 이러한 파트너십이 지역 수준에서 어떻게 작용할 수 있는지를 분석함.
- 전문 AI 조달 프레임워크 또는 동적 구매 시스템에 따르면 후속 계약에 적용되는 약관을 규정하고 윤리적 요구사항을 포함하는 일련의 사전적 정의 기준에 따라 공급업체를 평가할 수 있음. 공공조달청의 동적 AI 구매시스템은 이런 새로운 접근방식의 첫 번째 사례에 해당함.

2. 공고

어떤 시장 접근 경로를 선택하든 AI 기술이 빠르게 발전하고 있으며 신 기술과 제품이 끊임없이 출시되고 있다는 점을 유념해야 한다. 입찰 공고에서 산출물 기반 요구사항을 사용하고 현재 직면하고 있는 문제와 기회를 설명하는 데 초점을 맞춘다. 이는 공급업체들로 하여금 어떤 기술이 당신의 요구사항에 가장 적합한지 판단할 수 있게끔 해줄 것이다.

요구사항 초안 작성

요구사항 초안은 혁신을 주도하는 한편 효과적이고 책임 있고 윤리적인 AI 기술 배치를 위한 기초를 수립할 수 있음.

산출물 기반 요구사항을 사용하여, 공급업체가 요구사항에 어떻게 대응할 것인지 제안하게 할 것. 사용자 요구 및 요구 성능에 의해 뒷받침되는 충분히 상세한 문제 진술 초안을 작성할 것.

AI 요구사항 초안을 작성할 때 고려해야 할 주요 사항은 다음과 같음.

문제 진술부터 시작할 것

제한 사항과 추가적인 기능 요구사항을 포함하여 어떤 과제를 해결하려는 목표인지 명확히 진술할 것. AI가 당면 과제와 관련이 있다고 생각하는 이유와 대안적 솔루션에 열려 있는 이유를 설명할 것.

데이터 전략 및 요구 사항을 강조할 것

데이터 발굴에 기반하여 AI 시스템이 귀 기관의 현재 데이터 전략 및 관행에 어떻게 부합하는지 설명할 것. 가능한 경우 개인정보 보호 영향 평가를 참조하고 데이터 평가 결과, 데이터 요구사항 및 데이터 거버넌스 접근 방식에 대한 세부 사항을 추가할 것.

데이터 품질, 편향성 및 제한에 주목할 것

데이터 발굴 과정에서 이해하게 된 점을 활용하여 입찰 공고에 데이터의 드러난 한계점을 강조하고 공급업체에게 이러한 단점을 해결하는 전략의 설명을 요청할 것. 귀 기관이 놓쳤을 수 있는 관련 제한 사항을 해결하기 위한 계획을 세우고 공급업체에게 이 문제를 완화하기 위한 전략을 요청할 것.

귀 기관이 공급업체의 AI 접근 방식을 이해해야 할 필요성을 강조할 것

공급업체가 변수를 선택하는 방법과 모델이 기반하고 있는 AI 기법(예: 감독형, 비감독형 또는 강화학습형 등) 등 알고리즘과 모델에 대한 정보를 알 수 있는 평가 질의 초안을 작성할 것. 또한 모델의 한계를 설정하기 위해 노력할 것. 공급업체가 프로젝트에 결부된 계획 하에 알고리즘을 훈련시키는 데 사용한 데이터의 출처와 특성에 대해 명확성을 추구할 것. 알고리즘에 대한 독립적 감사도 요구사항에 포함할 것을 고려할 수 있음. 평가 기준이 이러한 지점들을 적절하게 평가하는지 확인할 것.

AI 시스템 '블랙박스' 및 공급업체에 종속되지 않는 전략을 고려할 것

'블랙박스'인 알고리즘에 의존하지 말 것. 입찰 공고에서 AI 개발에 대해 '설명 가능한 접근법'의 필요성(AI 시스템의 의사결정 과정을 이해할 수 있는 범위)을 강조할 것. 고도로 '설명 가능한' AI 시스템의 산출물은 귀 기관 팀은 물론 다른 공급업체에 의해 해석될 수 있을 것임. 이를 통해 향후 다른 공급업체와 협력하여 초기 AI 시스템을 유지하거나 개선할 수 있는 가능성이 높아져 공급업체에 종속될 위험을 제한함. 조달 문서에서 이 문제를 다루는 것을 고려할 것. 개방형 표준, 저작권료 면제 허가 계약(royalty-free licensing agreements) 및 퍼블릭도메인 공개 약관(public domain publication terms)의 채택이 모범 관행이 될 수 있음.

<데이터 윤리 프레임워크> 원칙 6 "투명하게 작업하고 책무를 이행해야 한다"를 적용하고 독립 감사와 같은 다른 수단의 사용을 고려할 것. 하십시오. 이 주제에 대한 자세한 정보는 물론 내용은 "설명 가능한 AI에 대한 왕립학회의 정책 브리핑" 참조.

지적재산권의 중요성을 다음과 같이 명시할 것

공급업체가 자신의 지적재산권(IP)과 상업적 이익을 보호하기 위해 솔루션 내부 작업에 대한 세부사항을 공개하는 것을 원치 않을 수 있음. AI 시스템의 설계와 배치 과정에서 새로운 알고리즘이 개발되거나 기존 알고리즘을 맞춤(예를 들어 데이터를 통한 알고리즘 재훈련)할 가능성이 높음. 귀 기관 또는 공급업체가 새로운 지적 재산을 소유해야 하는지 여부와 누가 생성된 IP를 가장 잘 활용할 수 있을지를 고려할 것. 협약은 상호 이익 및 공정함을 보장할 것.

제10차 정부 설계 원칙을 고려하여 귀 기관의 작업이 공개되고 다른 사람이 재사용할 수 있도록 보장 할 것.

관련 기술 또는 서비스 통합에 대해 언급할 것

귀 기관 데이터 및 기술 설계자와 협력하여 AI 시스템과 통합해야 할 관련 기술 또는 서비스를 정의할 것. 공급업체가 입찰 공고에 응할 때 알아야 할 특수 설계 기준이 있을 수 있음.

지속적인 지원 및 유지보수에 대한 귀 기관의 요구사항을 고려할 것

운영부서 또는 서비스부서의 직원은 AI 시스템의 사용 방법과 그 산출물에 대한 조치를 이해할 수 있도록 AI 시스템에 대해 충분한 지식을 갖추었거나 교육훈련을 받아야 함. 귀 기관은 교육훈련 및 지식 이전의 중요성을 강조하여 귀 기관의 팀이 계약 기간 동안 숙련도가 향상되고 공급업체가 시행한 솔루션에 대해 심층적으로 이해할 수 있도록 보장하고, 비전문가인 사용자가 어떻게 지원받을 수 있을지 고려할 수 있음.

최초 계약 기간을 넘어 요구될 수 있는 지속 지원에 대한 요구사항, 호스팅 또는 추가 개발 사항을 고려할 것. 서비스 지원 및 유지보수를 담당하는 기관내 팀이 있는 경우, 해당 팀이 공급업체와 상담했는지 확인할 것.

책임 및 위험성에 대한 고려사항을 추가할 것

위험성은 이를 가장 잘 관리할 수 있는 측에 할당할 것. 알맞은 위험성 배분은 서비스의 장기적 실행 가능성뿐만 아니라 최고 가치 달성을 위해서도 중요함. 특정 영역에 대한 책임은 특정 부서, 특히 AI 기반 솔루션의 사용 및 적용, 데이터 접근 및 이전과 관련된 부서에 속할 것임. 또한 기술, 보안 및 품질 보장에 관한 영역 등에서는 공급업체에게 책임을 물을 필요가 있을 수 있음. 이러한 고려 사항들에 대하여 입찰 공고에서 강조할 것. 위험 할당에 대한 자세한 내용은 <아웃소싱 플레이북> 8장 참조.

또한 신청서는 인가되지 않은 것으로 의심되는 행위를 기관 내부 또는 외부의 관련 기관에 쉽게 신고할 수 있는 방식을 포함해야 함.

귀 기관 팀의 경험을 활용하여 초안 작성 절차를 지원하고, 사용자 및 이해관계자와 협력하여 포함되어야 하는 평가의 핵심 영역을 설정할 것. 모든 요구사항은 투명해야 하며, 중소기업 및 사회적 기업과 같은 특정 유형의 공급업체 또는 영국이 조달 의무와 함께 무역 협정을 체결한 국가의 공급업체들을 차별해서는 안 됨.

더 알고 싶다면

<아웃소싱 플레이북>에 설명된 상업적 모범 관행을 따르고 <기술 규칙> 및 <정부 설계 원칙>을 준수해야 한다.

3. 선정, 평가 및 낙찰

선정 및 평가 단계에서는 요구사항에 대한 공급업체의 반응을 고려해야 한다. 다학제 팀의 경험을 활용하여 평가 절차를 지원하고 평가를 수행할 수 있는 광범위한 전문지식을 확보해야 한다.

견고성 실행 원칙의 주요 내용에 따르면, 공급업체가 AI 기반 솔루션을 제공할 때 시연을 요청하고 입찰자를 평가할 때 이를 추구해 볼 수 있다. 견고성 실행 원칙에는 다음이 포함될 수 있지만 이에 국한되지는 않는다.

- 사내 윤리적 AI 접근 방식을 수립하고, AI 기반 솔루션의 설계, 개발 및 배치 방법에 대한 사례를 제시
- 알고리즘 산출물에 대한 책무성을 보장하는 절차
- 불공정하게 차별할 수 있는 산출물의 회피
- 재현가능성을 위한 설계
- 다양한 조건에서 이루어지는 모델 테스트
- 허용 가능한 모델 성능에 대한 정의
- 견고하고 비례적인 보안의 제공

평가 절차의 일환으로 AI 시스템을 개발하고 배치할 팀의 전문 기술, 자격, 다양성 등도 검토한다. 이는 또한 시스템의 불공정한 편견을 예측하거나 탐지하는 데 도움이 될 수 있다.

공급업체의 반응은 공급업체의 일반적인 접근 방식을 나타내는 지표라는 점에서 유의해야 한다. 완전하게 상세하고 확고한 계획을 기대하기는 어려운데, 이는 AI 개발이 반복적인 과정이고 프로젝트가 진행됨에 따라 시스템이 예외없이 변화하고 진화할 것이기 때문이다.

더 알고 싶다면

세계경제포럼이 인공지능실과 협력 개발한 참고자료들도 AI 입찰자의 반응을 평가하는 사례가 될 수 있으며, 평가 단계에서 추가할 고려사항에 대한 아이디어를 제공한다. 마젠타북은 정부의 모니터링 및 평가 전략에 대한 일반적인 지침을 제공할 수 있다.

4. 계약 이행 및 지속적인 관리

다른 모든 계약과 마찬가지로, 새로운 공급업체가 <아웃소싱 플레이북> 및 계약 관리를 위한 모범 관행에 따라 참여할 수 있도록 시간과 주의를 기울여야 한다.

AI 시스템은 생애주기 전체에 걸쳐 지속적인 지원이 필요할 수 있다. 지원 격차에 따른 영향력을 수용하거나 외부 전문지식을 채택하는 것은 둘 다 비용이 든다. AI 기반 솔루션을 구매할 때 이 점을 고려해야 한다.

절차 기반 거버넌스 및 감사가능성

<AI 윤리 및 안전 이해 지침>에 제시된 절차 기반 거버넌스 프레임워크의 실시를 고려할 것. 이는 규범, 가치, 원칙에 대한 고지 절차와 프로젝트 업무흐름을 정의하는 실행규약을 통합할 수 있는 근거가 됨. 앨런 튜링 연구소는 이를 '절차 기반 거버넌스(Process-Based Governance) 프레임워크'라고 명명하였으며 이는 귀 기관 팀에 다음의 개요를 제공할 수 있음.

- 각 거버넌스 조치에 관련된 해당 팀원 및 역할
- 거버넌스 목표를 달성하기 위해 개입과 맞춤형 고려사항이 필요한 업무 흐름 단계
- 평가, 후속 조치, 재점검 및 지속적인 감독에 대한 명시적 시간표
- 명확하고 잘 정의된 실행규약으로 단대단 감사가능성을 지원하는 활동 기록(log)과 집행 체계

프로젝트 생애주기의 모델링, 교육훈련, 테스트, 검증 및 구현 단계에

걸쳐 데이터를 수집하는 과정이 기록되도록 구현하여 단대단 감사가 가능해야 함. 해석 가능하고 정당한 AI를 구현하기 위하여 이러한 기록은 다양한 사용자를 염두에 둔 가변적 접근성 및 정보 표출이 이루어져야 함.

모델 테스트

모델의 정확성을 유지하기 위해서는 지속적인 모델 테스트가 필요함. 부정확한 모델은 국민에게 부정적인 영향을 미치는 잘못된 결정을 초래할 수 있음. 따라서 일단 배치한 모델의 유효성을 모니터링할 방법을 공급업체와 함께 수립할 것. 국가 사이버 보안 센터에서 발간한 사이버 보안을 위한 지능적 도구 평가 지침은 또한 이러한 고려사항의 중요성을 강조하였음.

지식 이전 및 교육훈련

프로젝트가 완수되면 내부적으로 팀들이 이 도구를 자체적으로 적절하게 사용할 수 있도록 지식 이전 계획의 완전성 및 논리를 평가할 것. 이를 보장할 수 없는 경우 채용 및 유지보수 또는 추가 유지보수 계약을 통해 사내 역량을 확립하도록 할 것.

운영부서 또는 서비스부서의 직원은 AI 시스템의 사용 방법과 그 산출물에 대한 조치를 이해할 수 있도록 AI 시스템에 대해 충분히 지식을 갖추었거나 교육훈련을 받아야 함. AI 애플리케이션 오남용을 방지하기 위해 직원 교육훈련 및 지원에 대한 요구 문제를 AI 공급업체와 해결할 것. 신청서는 인가되지 않은 것으로 의심되는 행위를 기관 내부 또는 외부의 관련 기관에 쉽게 신고할 수 있는 방식을 포함해야 함.

수명 종료

AI 시스템 및 데이터의 수명을 종료시키는 절차를 어찌해야 할지 고려해 볼 것. 이 지침을 효과적으로 적용하기 위해서는 데이터 청소 및 수집에 대해 감사 가능한 방법론이 핵심임. 계약 종료의 역할과 절차를 정의

하는 것이 계약기관 및 공급업체 모두에게 중요함. 그러한 고려사항이 계약에 포함되어 있는지 확인하고, 광범위한 절차의 일부로서 계약관리 절차가 AI 시스템의 수명 종료를 적절하게 뒷받침할수 있을 정도로 충분히 견고한지 검사할 것. ▢

공교육에 적용되는 인공지능 알고리즘의 공공성 확보방안 연구

인쇄일 : 2021년 3월

발행일 : 2021년 3월

발행처 : 서울특별시교육청 교육혁신과

인쇄처 : 한울타리 (TEL : 02) 924-9642)

서울특별시교육청