공공분야 AI 시스템 도입과 개인정보영향평가

이은우(법무법인 지향, 정보인권연구소 이사)

다양한 제도적 접근

- 정보공개
- 개인정보영향평가
- 알고리즘 영향평가
- 인권영향평가
- 기타 거버넌스
- 행정절차법(공청회, 청문회)

GDPR의 개인정보보호 영향평가

- 다양한 기능으로 제도 설계
  - GDPR은 신기술의 도입이나, AI 시스템, 자동의사결정이나 프로파일링의 도입 등에 대응하여 개인정보보호 영향평가를 도입했는데, 다양한 기능을 수행할 수 있도록 제도 설계
- 새로운 개인정보처리의 위험성 파악과 영향의 이해
  - GDPR은 개인정보보호 영향평가를 새로운 개인정보처리 기술이나 처리절차에 대해서 그 영향과 위험 등을 미리 파악하고 이해할 수 있게 하는 기능과 역할을 하도록 함.
  - 특히 신기술의 영역이나, 자동의사결정 시스템, 대규모 시스템이나 프로파일링이 포함된 시스템 등에서 매우 중요한 역할을 할 것으로 기대
  - 이와 같은 기술이나 시스템으로 인한 위험이나 영향을 파악하고, 이해하는 것을 분명하게 하기 위하여 GDPR은 이런 내용을 정리하여 기술하는 것을 개인정보보호 영향평가에 포함.
  - 개인정보처리에 대한 이해를 분명하게 기술하게 하고, 이를 의무화한다면 이는 투명성 보장에도 도움이 될 것

GDPR의 개인정보보호 영향평가

- 위험을 완화할 수 있는 적절한 수단의 채택과 비례성
  - 파악된 위험을 완화할 수 있는 적절한 수단을 채택하도록 하고, 그 수단이 위험에 대한 적절한 안전조치로서 법을 준수하는 수준의 비례성을 유지하는지를 평가하도록 함.
  - 이를 통해서 해당 개인정보 처리가 안전조치를 갖춘 것인지 평가가 이루어짐.
  - 개인정보보호 영향평가 절차를 통해서 이 점을 문서화하여 투명성 보장에도 도움
- 사전 예방과 처리의 라이프 사이클에 따른 대응
  - 개인정보처리를 하기 전에 수행하도록 함으로서 사전 예방의 수단이 되게 하는 것은 물론, 파악하지 못한 새로운 위험이 예견되거나, 발견되는 경우에는 개인정보 영향평가를 다시 수행하도록 하여 처리의 라이프 사이클에 따른 평가와 대응이 지속적으로 이루어질 수 있게 함
- 개인정보보호 영향평가를 통한 투명성, 책임성, 개인정보주체의 접근권 보장
  - 해당 개인정보처리의 위험성을 미리 파악, 이해하여 이를 기술하도록 하고, 위험을 완화하여 법이 요구하는 안전수준을 준수하는 개인정보 처리가 보장될 수 있다는 것을 평가, 기술하게 하여 투명성과 책임성, 개인정보주체의 접근권을 보장하는 수단으로 기능

GDPR DPIA 대상 – 고위험을 불러올 가능성이 있는 경우

- 고위험을 불러올 가능성이 있는 경우
  - 개인정보 처리의 성격과 범위, 상황, 목적을 참작하여, 특히 신기술을 사용하는 처리 유형이 개인의 권리와 자유에 중대한 위험을 초래할 것으로 예상되는 경우(*likely to result in a high risk to the rights and freedoms of natural persons*)를 요건으로 함
  - 구체적 예시
    - 프로파일링 등의 자동화된 처리에 근거한, 개인에 관한 개인적 측면을 체계적이고 광범위하게 평가하는 것으로 해당 평가에 근거한 결정이 해당 개인에게 법적 효력을 미치거나 이와 유사하게 개인에게 중대한 영향을 미치는 경우,
    - 특정범주의 개인정보에 대한 대규모 처리나 범죄경력 및 범죄 행위에 관련된 개인정보에 대한 처리,
    - 공개적으로 접근 가능한 지역에 대한 대규모의 체계적 모니터링을 예시

GDPR DPIA 대상 – 고위험을 불러올 가능성이 있는 경우

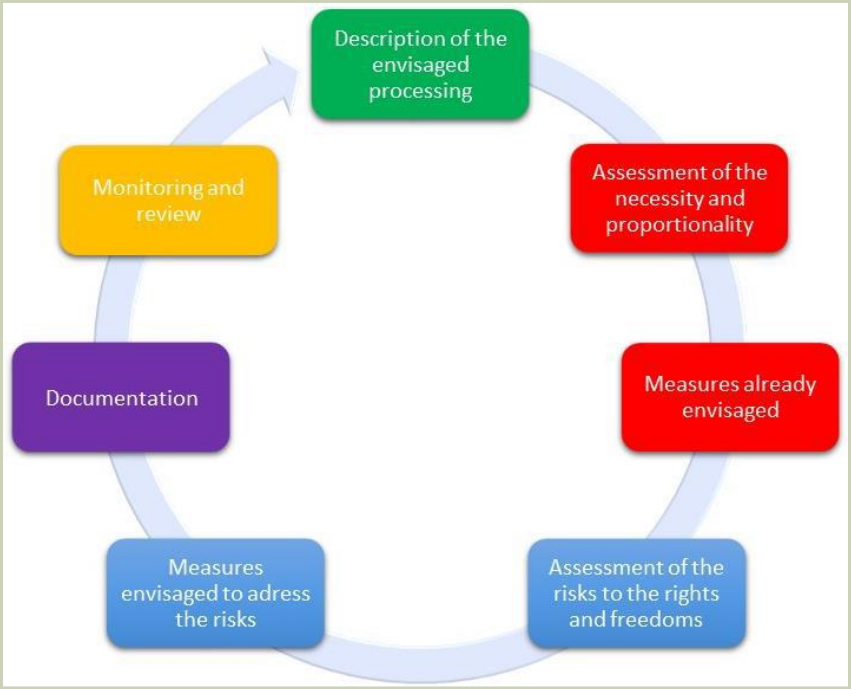
- DPIA 실시가 필요한 경우인지 판단하기 위한 기준에 대해서는 의견서
- 첫째, 평가 및 점수부여(scoring)의 경우
  - 정보주체의 직장에서의 실적, 경제사정, 건강, 개인의 선호나 관심, 신뢰성이나 행동, 위치, 동선에 대한 평가나 점수부여 등이 여기에 해당한다. 예를 들어 신용평가. 자금세탁, 반테러, 고객 건강평가 유전자 검사. 웹정보 기반 행동 프로파일, 마케팅 프로파일링이 여기에 해당한다.
- 둘째, 법적 효력 혹은 이와 유사한 중대한 영향을 미칠 수 있는 자동의사결정
- 셋째, 체계적인 모니터링이 이루어지는 경우
  - 체계적이라는 것은 시스템에 의해서 모니터링이 이루어지거나(occurring according to a system), 사전에 준비되거나 조직되어 있거나, 방법화되어 있는 것(pre-arranged, organised or methodical), 정보 수집의 일반적 계획의 일부로 이루어지는 경우(taking place as part of a general plan for data collection), 전략의 일환으로 이루어지는 경우(carried out as part of a strategy)를 의미. 네트워크를 통한 정보 수집도 마찬가지. 일반인이 접근할 수 있는 영역에 대한 체계적 모니터링도 이에 해당하는데, 이 경우는 정보주체가 인식할 수 없거나, 피할 수 없기 때문이라고 함. The WP29 interprets "*systematic*" as meaning one or more of the following (see the WP29 Guidelines on Data Protection Officer 16/EN WP 243):
- 넷째, 민감정보를 처리하는 경우

GDPR DPIA 대상 – 고위험을 불러올 가능성이 있는 경우

- 다섯째, 대규모 정보처리를 하는 경우
  - 이 경우 그 정보주체의 수나 관련되는 정보주체 중 처리되는 인구의 비율을 고려하여 판단하기도 하고, 처리되는 정보의 항목들의 양이나, 정보 항목의 범주의 다양성도 고려한다. 기간도 고려하고, 영구적인지도 고려. 처리 활동의 지역적 범위도 고려한다.
- 여섯째, 개인정보를 매칭하거나 개인정보 세트를 통합하는 경우
  - 이는 다른 목적으로 수집된 정보, 다른 처리자에 의해 수집된 정보가 매칭되거나 통합되는 것으로, 정보주체의 합리적인 기대를 초과하기 때문.
- 일곱째, 취약한 정보주체와 관련된 정보의 처리인 경우
  - 이는 쉽게 동의. 반대하기 어려운 점이나, 권리행사하기 어려운 점을 고려한 것. 예를 들어 아동. 피용자, 정신건강, 노인, 환자 등이 이에 해당.
- 여덟째, 정보의 혁신적 사용 혹은 정보처리 시 기술적·관리적 해결책을 적용하는 경우
  - 이는 새로운 기술을 활용하는 것이기 때문에 그로 인한 영향을 파악할 필요가 있기 때문이다.
- 아홉째, EU 역외로의 정보이전이 있는 경우.
- 열째, 정보처리 자체가 “정보주체로 하여금 본인의 권리행사를 막거나 서비스 혹은 계약을 이용하지 못하게 할 경우”
  - 이는 서비스 거부에 해당하기 때문에 권리나 자유에 큰 영향을 미치는 경우이다.

개인정보보호 영향평가의 시기

- 처리 이전
  - 개인정보보호 영향평가는 컨트롤러가 처리 이전에 '예정된 처리 작업이 개인정보보호에 미치는 영향에 대한 평가'를 시행하도록 함
- 위험 변화
  - 컨트롤러는 적어도 처리 작업으로 초래되는 위험에 변화가 있을 시에는 처리가 개인정보보호 영향평가에 따라 실시되는지를 평가하기 위한 검토를 시행해야 함
  - 위험의 변동이 생기는 경우에는 그 시점에 다시 영향평가를 해야 한다.
- 라이프사이클의 구조를 갖추고 영향평가를 시행해야 한다.
  - 단, 한 번의 평가로 유사한 중대한 위험을 초래하는 일련의 유사 처리 작업을 다룰 수 있다.



개인정보보호 영향평가의 내용

- GDPR이 규정한 개인정보보호 영향평가에 포함해야 할 최소한의 내용. 투명성 보장.
- 첫째, 예상되는 처리 작업 및 컨트롤러의 정당한 이익 등 개인정보 처리의 목적에 대한 체계적인 설명
  - 이는 개인정보 처리를 하기 전에 먼저 예상되는 처리 작업의 처리 목적을 명확하게 밝히도록 함으로서 목적의 명확화에 기여를 할 것이다.
- 둘째, 목적과 관련한 처리 작업의 필요성 및 비례성에 대한 평가.
  - 이처럼 먼저 목적을 명확하게 규정하고, 처리가 규정된 목적과 비교하여 필요성과 비례성을 충족하는지를 평가하도록 하는 것이다. 이는 최소수집의 원칙, 목적 명확화의 원칙을 평가하는 것이다.
- 셋째, 개인정보주체의 권리와 자유에 대한 위험성 평가
  - 이는 해당 개인정보의 처리로 인해서 개인정보주체의 권리와 자유에 어떤 위험이 초래될 수 있는 것인지. 그 영향은 무엇인지를 파악할 수 있게 하는 요소이다.
- 넷째, 개인정보주체와 기타 관련인의 권리 및 정당한 이익을 고려하여 개인정보의 보호를 보장하고 본 규정의 준수를 입증하기 위한 안전조치, 보안조치, 메커니즘 등 위험성 처리에 예상되는 조치.
  - 이는 위험이 발생하는 경우 그 위험을 제어하는 수단으로 안전조치, 보안조치, 메커니즘이 적절한지를 평가하는 것이다.
- GDPR은 승인된 행동강령을 따르기로 하였다면, 영향평가를 할 때 이를 준수하는지를 고려해야 함.

영향 평가과정에서 개인정보주체의 의견

- GDPR은 적절한 경우, 컨트롤러는 상업적 이익이나 공익의 보호 또는 처리 작업의 보안을 침해하지 않고, 예정된 처리에 대한 **개인정보주체 또는 그 대리인의 의견을 구해야** 한다고 규정하고 있다.

개인정보보호 영향평가 후의 사전협의

- GDPR은 개인정보보호 감독기구로 사전자문을 구하도록 하는 절차를 두어서 예정된 개인정보 처리가 초래하는 위험을 파악하고, 그에 대한 위험 억제 수단을 적절하게 갖추지 못한 경우에는 개인정보보호 감독기관이 조치를 취할 수 있도록 하였다.

- **개인정보보호 감독기관에 사전자문을 얻어야 할 경우**
  - GDPR은 개인정보보호 영향평가를 시행한 결과 처리가 고위험의 결과를 초래하는 경우로서, 개인정보 관리자가 그 위험을 완화하기 위해 취한 조치가 부재하는 경우에는 개인정보 처리 전에 개인정보 감독기구의 사전 자문을 구해야 한다는 규정을 두고 있다. 즉, 영향평가에서 해당 개인정보 처리가 고위험을 초래할 가능성이 있는데, 개인정보 관리자가 그 위험을 상쇄시킬 수 있는 안전조치를 취할 수 없거나, 안전조치를 취하였지만 그 안전조치가 위험을 없앨 수 있는 적절한 수단으로 보기 어려운 경우에는 개인정보 처리 전에 개인정보 보호 감독기구의 사전 자문을 얻도록 한 것이다.
- **제공할 자료**
  - 개인정보 관리자는 이때 다음과 같은 자료를 감독기구에게 제공해야 한다.
  - ① 가능한 경우, 처리에 관여하는 컨트롤러, 공동 컨트롤러 및 프로세서의 개별 책임, 특히 사업체집단 내의 처리에 대한 책임, ② 예정된 처리의 목적 및 방법, ③ 본 규정에 따라 개인정보주체의 권리와 자유를 보호하기 위해 제공되는 조치 및 안전조치, ④ 가능한 경우, DPO의 상세 연락처, ⑤ 개인정보보호 영향평가, ⑥ 감독기관이 요청한 기타 정보,
- **필요한 조치**
  - 이때 감독기관이 해당 예정된 처리에 대해서 개인정보 관리자가 위험을 충분히 파악하지 못하였거나, 해당 위험을 완화하지 못한 경우 등 해당 처리가 개인정보보호 규정을 위반할 것으로 의견을 제시하는 경우에는 8주 이내에 서면 형식의 권고를 제공해야 한다. 아울러 이때는 개인정보처리를 하기 전이라도 처리의 정지 등의 조치를 취할 수 있다.

개인정보영향평가(개인정보보호법)

개인정보 영향평가 대상

- 공공기관으로 한정

개인정보 영향평가 의무

- 개인정보파일에서 처리하는 개인정보의 양을 기준으로 대상 결정

3가지 경우

- 공공기관에서 100만명 이상의 정보주체에 대한 개인정보파일을 구축, 운용, 변경하려는 경우
- 공공기관이 개인정보파일을 연계하려는 경우 : 50만명 이상
- 공공기관이 민감정보나 고유식별정보 처리를 하는 경우 : 5만명 이상

(고시)개인정보 영향평가의 평가영역 및 평가분야는 인공지능 알고리즘

평가 영역	평가 분야	세부 분야	
Ⅰ. 대상기관 개인정보 보호 관리체계	1. 개인정보보호 조직	개인정보보호책임자의 지정	
		개인정보보호책임자 역할수행	
	2. 개인정보보호 계획	내부관리계획 수립	
		개인정보보호 연간계획 수립	
	3. 개인정보 침해대응	침해사고 신고 방법 안내	
		유출사고 대응	
	4.정보주체권리보장	정보주체 권리보장 절차 수립	
		정보주체 권리보장 방법 안내	
Ⅱ.대상시스템의 개인정보 보호 관리체계	5.개인정보취급자 관리	개인정보취급자 지정	
		개인정보취급자 관리・감독	
	6. 개인정보파일 관리	개인정보파일대장 관리	
		개인정보파일 등록	
	7. 개인정보처리방침	개인정보처리방침의 공개	
		개인정보처리방침의 작성	

개인정보 영향평가의 평가영역 및 평가분야

III. 개인정보 처리단계별 보호조치	8. 수집	개인정보 수집의 적합성	
		동의 받는 방법의 적절성	
	9. 보유	보유기간 산정	
	10.이용·제공	개인정보 제공의 적합성	
		목적 외 이용・제공 제한	
		제공시 안전성 확보	
	11. 위탁	위탁사실 공개	
		위탁 계약	
		수탁사 관리・감독	
	12. 파기	파기 계획 수립	
		분리보관 계획 수립	
		파기대장 작성	
	13. 접근권한 관리	계정 관리	
		인증 관리	
		권한 관리	

개인정보 영향평가의 평가영역 및 평가분야

IV. 대상시스템의 기술적 보호 조치	14. 접근통제	접근통제 조치	
		인터넷 홈페이지 보호조치	
		업무용 모바일기기 보호조치	
	15. 개인정보의 암호화	저장시 암호화	
		전송시 암호화	
	16. 접속기록의 보관 및 점검	접속기록 보관	
		접속기록 점검	
		접속기록 보관 및 백업	
	17. 악성프로그램 등 방지	백신 설치 및 운영	
		보안업데이트 적용	
	18. 물리적 접근방지	출입통제 절차 수립	
		반출·입 통제 절차 수립	
	19. 개인정보의 파기	안전한 파기	
	20. 기타 기술적 보호조치	개발 환경 통제	
		개인정보처리화면 보안	
		출력시 보호조치	
	21.개인정보처리구역 보호	보호구역지정	

개인정보 영향평가의 평가영역 및 평가분야

V. 특정 IT기술 활용시 개인정보 보호	22. CCTV	CCTV 설치시 의견수렴	
		CCTV 설치 안내	
		CCTV 사용 제한	
		CCTV 설치 및 관리에 대한 위탁	
	23. RFID	RFID 이용자 안내	
		RFID 태그부착 및 제거	
	24. 바이오정보	원본정보 보관시 보호조치	
	25. 위치정보	개인위치정보 수집 동의	
		개인위치정보 제공시 안내사항	

[별표4]에 명시되지 않은 특화된 IT기술을 적용하는 경우에는 해당 기술이 개인정보 보호에 미치는 영향에 대한 평가항목을 개발하여 영향평가 시 반영하여야 한다.

ICO AI 감사 프레임웍(AI AUDITING FRAMEWORK)

- ICO의 AI 감사 프레임웍(AI Auditing Framework)
  - AI 감사 프레임웍 가이드 초안(2020. 2. 19.)

1. Accountability and governance

- data protection impact assessments (DPIAs),
- controller / processor responsibilities,
- assessing and justifying trade-offs.

2. Fair, lawful and transparent processing in AI systems

- lawful bases for processing personal data in AI systems
- assessing and improving AI system performance
- mitigating potential discrimination to ensure fair processing.

3. Data minimisation and security in AI systems

4. The exercise of individual rights

- how you can ensure meaningful human input in non-automated or partly-automated decisions,
- meaningful human review of solely automated decisions.

AI 시스템에 대한 DPIA(ICO 초안)

1. 처리에 대한 체계적인 기술(Systematic description of the Processing)

- DPIA에는 처리 활동에 대한 체계적인 기술이 포함되어야 함. 여기에는 개인에게 영향을 미칠 수 있는 AI 처리와 자동 결정이 이루어질 때, 그 데이터 흐름과 단계들이 포함되어야 함. 모든 변수의 적절성과 오차의 한계를 설명할 수 있어야 한다.
- 자동화된 결정이 인간의 개입이나 검토(human intervention or review)를 받아야 하는 경우에는 이를 위한 절차가 실질적으로 의미있도록 보장해야 하고, 번복될 수 있는 결정들이 구체화되어 있어야 한다.
- 이 절차를 생략해야 할 합리적 이유가 없는 한 해당 조직은 DPIA가 이루어지는 당해 처리에 관한 개인들 또는 그 대리인들의 입장을 확인하고 문서화해야 한다. 따라서 당해 당사자들이 확인할 수 있도록 처리들을 설명할 수 있는 것이 중요하다.
- 그러나 복잡한 AI 시스템의 처리 활동을 기술하는 것은 어려울 수 있다. 따라서 평가를 위해 두 가지 버전의 설명을 만드는 것이 적합할 수 있다. 하나는 전문가들을 위한 기술적 기술로 표현하는 것이고, 두 번째는 처리에 대해서 좀 더 높은 수준의 기술을 하고, 투입되는 개인정보들과 개인에게 영향을 미치는 결과물의 관계가 어떻게 되는지에 대한 논리의 설명을 담고 있는 것이다.
- DPIA에는 정보 관리자와 모든 처리자들의 역할과 의무를 설정해야 한다.AI system이 전부나 일부 외부 공급자들에게 아웃소싱되는 경우에는 양 조직은 공동 관리자가 되는 것인지 여부도 평가해야 한다.이 경우 양자는 DPIA에서도 협력이 필요하다.
- 개인정보 처리자가 활용되는 경우에는 DPIA에 개인정보 처리의 다소 기술적인 속성들은 해당 개인정보 처리자가 제공한 정보를 활용하여 설명될 수 있다. 예를 들어 개인정보 처리자의 매뉴얼의 흐름도 다이어그램 등. 이 경우 개인정보 관리자는 자신의 평가에서 처리자의 문헌을 많은 부분 복사해서 사용하는 것은 지양되어야 한다.

AI 시스템에 대한 DPIA(ICO 초안)

2. 필요성과 적정성 평가(Assessing necessity and proportionality)

- 특정한 목적, 합법적인 목적을 수행하기 위해 개인정보를 처리하려는 AI 시스템의 이행은 해당 시스템의 능력이 검증된 것(proven ability of that system)에 의하여 이루어져야지, 그 기술의 가능성(availability of the technology)에 의하여 이루어져서는 안된다. DPIA의 필요성 평가에 의하여 By assessing necessity in a DPIA, an organisation can evidence that these purposes couldn't be accomplished in another reasonable way.
- DPIA를 수행함에 의하여 조직은 AI 시스템에 의한 개인정보 처리가 적정성이 인정되는 활동임을 증명할 수 있다.적정성을 평가할 때 조직의 이익과 개인의 권리와 자유와 비교 산출되어야 한다. AI 시스템과 관련하여 조직은 알고리즘과 사용되는 데이터의 부정확성 또는 편견으로 인해 정보주체에게 미칠 수 있는 모든 침해를 생각할 필요가 있다.
- DPIA의 적정성 평가요소에서 조직은 정보주체가 AI 시스템에 의하여 정보가 처리될 것이라고 합리적으로 예측할 수 있을지 여부를 평가해야 한다.만약 AI system이 인간의 의사결정을 보완하거나 대체하는 경우 DPIA에 그 프로젝트가 인간과 알고리즘의 정확성에 대하여 각 측면을 상호 비교하여 사용을 정당화할 수 있는 수단이 무엇인지를 문서화해야 한다.
- 조직은 예를 들어 정확성과 데이터 최소화의 상호 상쇄효과와 같은 장단점과 상호 상쇄되는 모든 효과(trade-offs)에 대해서도 기술하고, 그에 대한 방법론과 논거를 문서화하여야 한다.

AI 시스템에 대한 DPIA(ICO 초안)

3.주체에 가해질 위험 확인 Identifying risks to rights and freedoms

- AI 시스템을 개발하는 과정과 전개하는 과정에서 개인정보 사용이 개인의 프라이버시와 개인정보보호에 대한 권리에 위험을 초래하는 것이어서는 안된다.
- 개인정보에서 역사적인 패턴으로부터 머신 러닝 시스템이 차별을 재생산할 수 있는데, 이는 법률의 평등권 침해이거나 차별일 수 있다. AI 시스템이 창작자의 개인정보 보호에 대한 분석에 입각하여 특정 콘텐츠의 출판을 중단하는 경우, 이는 그들의 표현의 자유에 영향을 줄 수도 있다. 이런 맥락에서 개인정보 관리자는 개인정보 보호 외의 다른 관련되는 법적인 구조를 고려해야 한다.
- DPIA 절차는 조직에서 관련되는 위험들을 객관적으로 식별하는 것에 도움이 될 것이다. 각각의 위험에 대해 점수나 레벨이 부여되어야 하고, 정보주체에게 미치는 영향의 중요성과 발생가능성과 관련하여 측정되어야 한다..

AI 시스템에 대한 DPIA(ICO 초안)

4.위험에 대해 대처할 수단 Measures to address the risks

- (가장 초기 단계부터, 프로젝트팀과의 열린 소통 채널)It is important that data protection officers and other information governance professionals are involved in AI projects from the **earliest stages**. Clear and open channels of communication must be **established between them and the project teams**. This will ensure that risks can be identified and addressed early in the AI lifecycle.
- (개인정보보호는 처음부터)Data protection should not be an afterthought, and a DPO's professional opinion should not come as a **surprise at the eleventh hour**.
- (충분히 훈련된 인력) A DPIA can be used to document the safeguards put in place to ensure the individuals responsible for the development, testing, validation, deployment, and monitoring of AI systems are **adequately trained** and have an appreciation for the data protection implications of the processing.
- (인간 오류, 적절한 훈련)Organisational measures to ensure that **appropriate training** is in place to mitigate risks associated with **human error** can also be evidenced in a DPIA. Along with the technical measures designed to reduce risks to the security and accuracy of an AI system.
- (억제된 위험과 남은 위험)Once measures have been introduced to mitigate the risks identified, the DPIA should document the **residual levels of risk posed by the processing**. These must be referred to the ICO for prior consultation if they remain high.

AI 시스템과 개인정보보호영향평가

5. 지속적인 문서화 A 'living' document
   - (생문서, 지속적인 정기적 리뷰와 재평가)While any DPIA must be carried out before the processing of personal data begins, they should be considered a **'live' document**. This means they are subject to **regular review or re-assessment** should the nature, scope, context or purpose of the processing alter for any reason.
   - (인구구성 변화, 행동 변화)For instance, depending on the deployment, it could be that the **demographics of the target population may shift**, or that people adjust their **behaviour over time** in response to the processing itself.

신뢰가능한 AI 구현을 위한 원칙(OECD)

- OECD는 2019년 5월 ‘OECD AI 권고안’을 회원국 만장일치로 공식 채택,
  - 신뢰가능 AI 구현을 위한 5가지 원칙과 신뢰가능 AI 시스템을 정의

OECD의 신뢰가능 AI의 원칙

- OECD는 'OECD AI 권고안(2019.5)'을 통해 다음과 같이 신뢰가능 AI 구현을 위한 5가지 원칙과 신뢰 가능 AI 시스템을 정의 원칙
- 원칙 1. 포용 성장, 지속가능 발전과 복지 증진
  - 모든 AI 이해관계자는 인류의 포용 성장, 지속 가능 발전 및 복지 증진을 위해 신뢰가능한 AI의 구현 에 힘써야 함
    - AI 이해관계자는 인간의 능력과 창의력을 향상시키고 소수집단 포용을 진전시키는 방향으로 AI의 구현에 힘써야 함
    - 아울러 성차별등 사회적 불평등을 감소시키고, 자연환경을 보호하는 방향으로 AI의 구현에 힘써야 함
- 원칙 2. 인간중심 가치와 공정성
  - AI 행위자는 AI 시스템 라이프사이클 전반에 걸쳐 법률, 인권, 민주적 가치, 공정 등 인간 중심 가치를 존중하고 지켜나가야 함
    - - AI 행위자는 자유, 존엄, 자치, 사생활 보호, 평등, 다양성, 공정성, 차별 금지, 노동권보장의 인간중심 가치를 지켜나가야 하며
    - - 이를 위해 AI 행위자는 인간중심 가치 실현을 위한 메카니즘과 인간이 최종의사결정에 개입할 수 있 는 안전장치를 마련해야 함

OECD의 신뢰가능 AI의 원칙

- 원칙 3. 투명성과 설명가능성
  - •AI 행위자는 사용자와 고객에게 AI 시스템에 대한 의미 있는, 최신 정보를 제공함으로써 그들과 적극 적으로 의사소통해야 함
    - - AI 행위자는 AI 시스템 활용 시점, AI 시스템 개발, 배치 및 운영 방식에 대해 투명하게 정보를 제공 할 수 있어야 함
    - - AI 행위자는 이해관계자에게 AI 시스템의 예측, 의사결정의 기저가 되는 핵심요인과 논리에 대하여 쉽게 설명할 수 있어야 함
- 원칙 4. 보안과 안전성
  - •AI 시스템은 전 수명주기에 걸쳐 어떠한 조건에서도 견고하게 작동되어야 하며 외부에 취약점이 노 출되지 않아야 함
    - - AI 행위자는 AI 시스템의 결과와 반응을 분석하기 위하여 데이터셋, 프로세스, 의사결정 과정 등을 추적할 수 있어야 함
    - - AI 행위자는 개인정보보호, 정보보안, 외부공격의 위험을 해소하기 위하여 체계적으로 위험관리방 법을 지속적으로 적용하여야 함
- 원칙 5. 책임성
  - •AI 행위자는 책임지고 AI 시스템 구현에 있어 위의 원칙을 실현하는 것과 AI 시스템이 올바르게 기능 할 수 있도록 노력해야 함
    - - 신뢰가능 AI 시스템 구현에 있어서 AI 행위자가 윤리적, 도덕적으로 행동할 수 있도록 안내하고 행 동강령 등을 명시함
    - - AI 행위자는 관련 문서를 제공하거나 경우에 따라서 감사를 받음으로 자신의 책임성에 입증할 수 있어야 함

OECD의 신뢰가능 AI의 원칙의 구현 실행가이드

공공기관 신뢰가능 AI의 구현 실행가이드

• 공공기관이 신뢰가능 AI 구현을 위한 원칙을 준수하기 위한 실행가이드

원칙	실행가이드
포용성장, 지속가능 발전, 복지증진	공공성의 확인 - AI 시스템의 기관 미션 연계성과 사회경제적 영향 평가
	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	AI 시스템에 대한 투명한 정보공개 AI에 관한 일반 정보, 개발/훈련/ 운영/활용의 방식에 관한 정보
	AI 시스템 결과에 대한 설명 요인, 데이터, 알고리즘 등 의사결정 요인과 전후 맥락 설명
보안 및 안전성	AI 시스템의 추적 가능성 보장 - 데이터 세트, 알고리즘, 프로세스 및 의사결정 관련 추적 가능성
	체계적인 위험관리 접근 - 가능한 위험 및 확률, 관리방안
책임성	AI 시스템 원칙의 실현 - 라이프사이클에서 발생한 의사결정과 행동 문서화

신뢰가능한 AI 구현 거버넌스 프레임워크

• 공공기관이 신뢰가능 AI 구현 실행에 관한 수행방법 가이드

방법	세부 지침
AI 거버넌스 구축과 운영	신뢰가능 AI 시스템 구현을 위한 협력과 소통 유도
	AI 시스템의 위험 평가 및 내부 관리
AI 시스템 위험평가와 의사결정 모델	AI 시스템의 위험의 확률과 심각성 평가
	AI 시스템 의사결정 프로세스에 인간의 개입 수준 결정
책임성 있고 우수한 데이터관리	데이터 이력관리
	데이터 품질관리
	데이터에 내재하고 있는 편향의 최소화
소비자 및 고객과 커뮤니케이션	AI 시스템 활용 관련 투명한 정보공개
	사람-AI 인터페이스 정보
	설명 정책 수립 및 추진

OECD의 신뢰가능 AI의 원칙과 영향평가(공공기관 신뢰가능 AI 구현 실용가이드)

AI System DPIA에 포함될 요소는?

1 원칙 1. 포용 성장, 지속가능 발전과 복지 증진

• 모든 AI 이해관계자는 국내외적으로 인류의 포용 성장, 지속 가능 발전 및 복지 증진을 위해 신뢰가능 AI의 구현에 힘써야 함

AI 원칙 실행가이드 \ AI 활동		기획	디자인	데이터	모델	검사실증	도입/설치/운영
		기획자	설계자	관리자	개발자	기획자/감리사	운영자
포용성장, 지속가능 발전, 복지증진	공공성의 확인 - AI 시스템의 기관 미션 연계성과 사회경제적 영향 평가	✓				✓	
	사회적 차별요소 배제 - 데이터, 모델로부터 성, 인종 등 차이로 인한 근원적 차별 배제		✓	✓	✓		✓

☐ AI 시스템의 공공성 확인

• AI 기획자/중간검수자는 AI 시스템 개발 전후로 다양한 이해관계자 협업과 토론을 통해 구현하려는 AI 시스템의 용도와 목적이 공공기관의 미션과 잘 부합하는지 확인해야 함

- 아울러 AI 시스템 구현을 통한 불평등, 환경오염, 인권차별 등의 부정적 영향이 일어날 가능성에 대한 논의도 함께 되어야 함

☐ AI 시스템의 사회적 차별요소 배제

• AI 시스템의 설계자/관리자/개발자는 데이터와 AI모델로 부터 성, 인종, 지역 등의 차이로 인한 근원적 차별을 배제해야 함

- AI 시스템의 운영자는 AI 시스템 활용결과, 사회적 차별요소가 감지되었을 때, 즉시 AI 행위자에게 즉시 관련된 모든 정보를 제공하여야 함

OECD의 신뢰가능 AI의 원칙과 영향평가(공공기관 신뢰가능 AI 구현 실용가이드)

2 원칙 2. 인간중심 가치와 공정성

• AI 행위자는 AI 시스템 라이프사이클 전반에 걸쳐 법률, 인권, 민주적 가치, 공정 등 인간 중심 가치를 존중하고 지켜나가야 함

AI 원칙 실행가이드 \ AI 활동		기획	디자인	데이터	모델	검사실증	도입/설치/운영
		기획자	설계자	관리자	개발자	기획자/감리사	운영자
인간중심 공정성	인간중심 가치와 공정성 촉진 - 인권영향평가, 인권실사, 윤리 행동강령, 품질인증 조치	✓				✓	✓
	인간중심 가치 내재화 - 적절한 안전장치 (Kill Switch, Human in the loop 등)		✓	✓	✓		✓

AI 시스템의 인간중심 가치와 공정성 촉진

• AI 시스템 기획자/중간검수자/운영자는 윤리적 행동강령을 마련하는 등 인간중심 가치와 공정성 촉진에 대한 약속을 분명히 하여야 함

- 아울러 상황에 따라 인권영향평가(참고 1) 및 인권실사(참고 2), 등의 조치를 취하여 잠재적 위험의 심각성, 인권에 미치는 영향을 식별해야 함

AI 시스템에 인간중심 가치 내재화

• AI 시스템의 설계자/관리자/개발자는 인간중심 가치와 공정성을 AI 시스템에 내재화하여 개발하는 것이 중요함

- 아울러 AI 시스템의 설계자/관리자/개발자는 상황에 따라 사람이 개입(human-in-the-loop)할 수 있고 감독할 수 있게 하는 등 적절한 안전장치를 갖춘 AI 시스템을 개발하여야 함(참고 3)

AI System DPIA에 포함될 요소는?

OECD의 신뢰가능 AI의 원칙과 영향평가(공공기관 신뢰가능 AI 구현 실용가이드)

3 원칙 3. 투명성과 설명가능성

• AI 행위자는 사용자와 고객에게 AI 시스템에 대한 의미 있는, 최신 정보를 제공함으로써 그들과 적극적으로 의사소통해야 함

AI 원칙 실행가이드 \ AI 활동		기획	디자인	데이터	모델	검사실증	도입/설치/운영
		기획자	설계자	관리자	개발자	기획자/감리사	운영자
투명성 설명가능성	AI 시스템에 대한 투명한 정보공개 AI에 관한 일반 정보, 개발/훈련/ 운영/ 활용의 방식에 관한 정보		✓	✓	✓		✓
	AI 시스템 결과에 대한 설명 - 요인, 데이터, 알고리즘 등 의사결정 요인과 전후 맥락 설명		✓	✓	✓		✓

❑ AI 시스템에 대한 투명한 정보 공개

• AI 시스템 설계자/개발자/운영자는 AI 시스템이 어떻게 개발되고, 훈련되며, 운영 및 활용되는지 방식을 이해관계자에게 공개

- 그러나 소스코드 및 데이터 세트는 영업비밀 또는 개인정보보호 등 지적 재산권의 보호 적용을 받을 수 있음

❑ AI 시스템 결과에 대한 설명

• AI 시스템 설계자/개발자/운영자는 AI 시스템이 의사결정에 어떻게 도달했는지에 관하여 이해관계자에게 설명할 필요가 있음

- 의사 결정의 주요 변수, 결정 요인, 데이터, 논리 또는 알고리즘을 명확하고 간단한 용어로 문맥에 따라 적절하게 제공
- 아울러 AI 시스템 특성상 유사한 환경임에도 불구하고 다른 결과를 생성한 이유를 설명할 필요가 있음

AI System DPIA에 포함될 요소는?

OECD의 신뢰가능 AI의 원칙과 영향평가(공공기관 신뢰가능 AI 구현 실용가이드)

4 원칙 4. 보안과 안전성

• AI 시스템은 전 수명주기에 걸쳐 어떠한 조건에서도 견고하게 작동되어야 하며 외부에 취약점이 노출되지 않아야 함

AI 원칙 실행가이드 \ AI 활동		기획	디자인	데이터	모델	검사실증	도입/설치/운영
		기획자	설계자	관리자	개발자	기획자/감리사	운영자
보안 및 안전성	AI 시스템의 추적 가능성 보장 - 데이터 세트, 알고리즘, 프로세스 및 의사결정 관련 추적 가능성		✓	✓	✓		✓
	체계적인 위험관리 접근 - 가능한 위험 및 확률, 관리방안	✓	✓	✓	✓	✓	✓

AI 시스템 추적 가능성 보장

• AI 시스템 설계자/개발자/운영자는 AI 시스템 라이프사이클 동안 이루어진 데이터 세트, 프로세스 및 의사 결정과 관련하여 추적이 가능하다는 것을 보장해야 함

- AI 시스템의 결과 및 문의에 대한 응답을 상황에 적합하고 최신 기술과 일치하도록 분석해야 함

AI 시스템 체계적인 위험관리 접근

• AI 행위자 모두는 각자의 역할, 상황 및 행동 능력에 따라 AI 시스템 수명주기의 각 단계마다 체계적인 위험 관리 접근 방식을 지속적으로 적용해야 함

- 위험관리에서 다루어야 할 위험으로 AI 시스템 고유의 위험, 개인 정보 보호, 디지털 보안, 데이터 편견 등이 있음

AI System DPIA에 포함될 요소는?

OECD의 신뢰가능 AI의 원칙과 영향평가(공공기관 신뢰가능 AI 구현 실용가이드)

AI System DPIA에 포함될 요소는?

5 원칙 5. 책임성

• AI 행위자는 책임지고 AI 시스템 구현에 있어 위의 원칙을 실현하는 것과 AI 시스템이 올바르게 기능할 수 있도록 노력해야 함

AI 원칙 실행가이드 \ AI 활동		기획	디자인	데이터	모델	검사실증	도입/설치/운영
		기획자	설계자	관리자	개발자	기획자/감리사	운영자
책임성	AI 시스템 원칙의 실현 - 라이프사이클에서 발생한 의사결정과 행동 및 조치 문서화	✔	✔	✔	✔	✔	✔

AI 시스템 원칙의 실현

• AI 행위자는 적용 가능한 윤리강령, 규정에 따라 앞서 언급한 신뢰가능 AI 구현을 위한 모든 원칙을 존중하고 그들의 행동과 의사 결정을 통해 이를 입증하여야 함

- 책임성 제고를 위해 AI 시스템 라이프사이클 동안 발생한 주요 결정 사항, 행동 및 조치를 문서화하여야 하고 기관내 윤리위원회, 규제위원회의 역할을 보장하고 적극 활용하여야 함

OECD의 신뢰가능 AI의 원칙과 영향평가(공공기관 신뢰가능 AI 구현 실용가이드)

4 신뢰가능 AI 구현 거버넌스 프레임워크

거버넌스 프레임워크 요약

• 거버넌스 프레임워크는 앞 장에서 설명한 공공기관이 신뢰가능 AI 구현을 위한 조치를 어떻게 수행할 것인가에 대한 가이드

방법	세부 지침
AI 거버넌스 구축과 운영	신뢰가능 AI 시스템 구현을 위한 협력과 소통 유도
	AI 시스템의 위험 평가 및 내부 관리
AI 시스템 위험평가와 의사결정 모델	AI 시스템의 위험의 확률과 심각성 평가
	AI 시스템 의사결정 프로세스에 인간의 개입 수준 결정
책임성 있고 우수한 데이터관리	데이터 이력관리
	데이터 품질관리
	데이터에 내재하고 있는 편향의 최소화
소비자 및 고객과 커뮤니케이션	AI 시스템 활용 관련 투명한 정보공개
	사람-AI 인터페이스 정보
	설명 정책 수립 및 추진

AI System DPIA에 포함될 요소는?

UK, A GUIDE TO USING ARTIFICIAL INTELLIGENCE IN THE PUBLIC SECTOR

- Assess, plan and manage artificial intelligence
  - Understanding artificial intelligence
  - Assessing if artificial intelligence is the right solution
  - Planning and preparing for artificial intelligence implementation
  - Managing your artificial intelligence project
- Using artificial intelligence ethically and safely
  - Understanding artificial intelligence ethics and safety