공개

의안번호	제 5 호
심 의 연 월 일	2026. 7. 23. (제 11 회)

보고사항

공공 AX
프라이버시 보호 안내서(안)

과학기술관계장관회의

제 출 자	개인정보보호위원회 위원장 송 경 희
제출 연월일	2026. 7. 23.

공공 AX 프라이버시 보호 안내서(안) (요약)

Ⅰ. 보고주문

○ 「공공 AX 프라이버시 보호 안내서(안)」를 별지와 같이 보고함

Ⅱ. 발간 배경 및 목적

○ **공공기관**은 **국민의 삶 전반**을 포괄하는 **방대한 데이터를 보유**, **대국민 서비스·정책 집행**을 통해 **강력한 영향력**을 행사

○ 이에, **공공 AX**의 **안전성**은 정부 **정책의 신뢰에 직결**, 프라이버시 보호는 **국민 신뢰 확보, 기본권 보장, 위험 통제**의 **핵심 수단**

○ 공공기관은 **개인정보보호 및 AI 전문인력**이 **부족**하여 **AI 데이터 처리방식**의 **구체적 이해**에 기반한 **개인정보보호에 애로**

※ (공공시스템 긴급 점검결과('26.3.)) 개인정보보호 **전담인력은 중앙 1.1명**, **기초지방정부 0.3명** 수준

⇒ **공공 AX 추진 시 개인정보 처리 이슈를 체계화**하고, **법적 기준과 안전조치, 관련 사례**를 제시하여 **공공 AX 적극 지원**

< 해외 유사 정책·가이드라인 사례>

미국	『**M-25-21: Accelerating Federal Use of AI through Innovation, Governance, and Public Trust**』(백악관 예산관리국(OMB), '25.4.) - 美 연방기관의 AI 활용, 거버넌스, 공공신뢰 확보를 위한 행정지침
EU	『**Orientations for ensuring data protection compliance when using Generative AI systems (Version 2)**』(유럽 개인정보감독기관(EDPS), '25.10.) - EU 기관·기구 등의 생성형 AI 활용시 개인정보보호 의무 이행 지원을 위한 가이드
영국	『**AI Playbook for the UK Government**』(영국 디지털서비스국(GDS), '25.2.) - 중앙정부, 공공기관 등이 AI를 안전하고 책임있게 활용할 수 있도록 지원하는 가이드

Ⅲ. 주요 내용

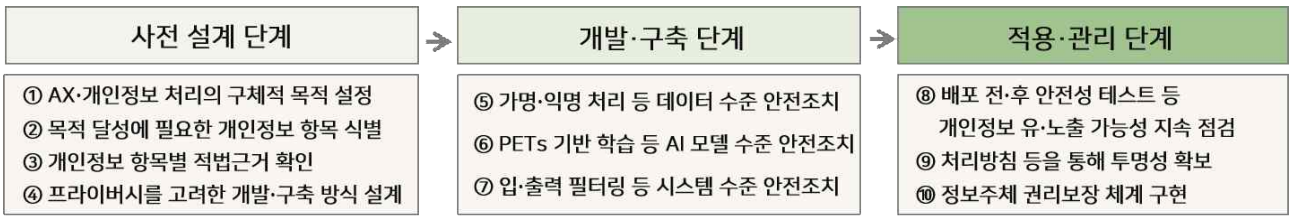
◈ **(기본 방향) 실무자가 쉽게 적용할 수 있도록 AX 추진 단계별 개요를 제시하면서, 개별 상황에 맞춰 참고할 수 있도록 AX 유형별 차등화된 방향 제시**

※ (AX 추진 단계) ① 사전 설계 단계 → ② 개발·구축 단계 → ③ 적용·관리 단계

※ (AX 유형 분류) ① 기초업무 지원 → ② 정보 연계·분석·추천 → ③ 선별·판단

1 공공 AX 단계별 검토

< 공공 AX 프라이버시 10대 핵심 점검 >



1 사전 설계 단계

○ **AX의 목적·대상**을 **명확히**하여 **목적 제한, 최소 수집** 등 **보호 원칙 준수** 기반을 마련하고, 개인정보 **수집·이용·제공의 적법근거 식별**

* AI 모델 및 시스템 개발·구축, AI 시스템 운영, AI 시스템 고도화 등 단계·목적별 구분

○ **AX의 목적, ICT·보안 역량**뿐만 아니라 활용되는 **개인정보의 성격, 프라이버시 리스크** 등을 종합적으로 검토하여 **개발·구축 방식 설계***

* △ 상용 서비스 활용, 공개모델 고도화, 자체개발 등 방식별 데이터 통제수준 비교, △ RAG, 추가학습(fine-tuning) 등 업무 특화 방식에 따른 프라이버시 리스크 검토

2 개발 · 구축 단계

○ **학습 데이터 구축·활용, 프롬프트 수집, RAG 연계, 출력 생성** 등 **처리 지점별**로 프라이버시 보호를 위한 **안전조치** 내재화

- **기본적 안전조치 外**에도 변화·발전 중인 **AI 특유의 프라이버시 리스크 유형, 경감방안**을 지속 검토하여 안전조치 **보완**

※ (예) △학습데이터 가명·익명 처리, 차분 프라이버시 등 개인정보보호강화기술(PETs) 활용, △개인정보 입력 제한·필터링, △RAG 접근권한 차등 부여, △개인정보 출력 필터링 등

3 시스템 적용 · 관리 단계

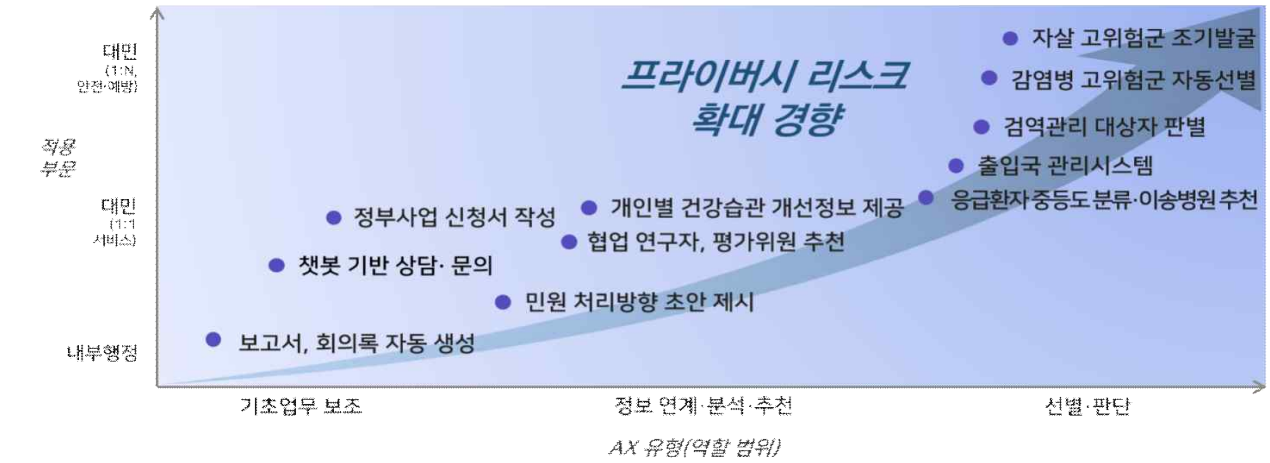
○ **배포 전·후 안전성 테스트**(red-teaming 등)를 통해 **개인정보 유·노출 가능성**을 **지속 점검**하고, **모니터링 및 사고대응** 프로토콜 마련

○ 개인정보 **열람, 정정·삭제, 처리정지** 등 **정보주체 권리 보장 체계** 구현*, **처리방침** 등을 통해 **수집 사실, 처리 목적** 등을 **투명하게 공개**

* 권리침해 신고 및 의견 제출 기능을 시스템에 탑재하고, 시간·비용·기술적 측면에서 실현가능한 범위에서 권리를 보장하기 위한 절차 마련

2 공공 AX 유형별 점검

※ AI 유형 및 적용 부문은 중첩 적용될 수 있으며 유연한 적용 필요



기초 업무 보조	정보 연계·분석·추천	선별·판단
안전한 활용 환경 조성, 기본적 안전조치에 중점	목적 외 이용 및 개인정보 추론 리스크 집중 관리	엄격한 적법성 검토, 자동화된 결정에 따른 리스크 관리
① 이용지침 안내 등 안전한 이용환경 조성 ② 오남용, 외부공격에 대한 시스템 안전성 확보 ③ AI 입·출력 데이터의 적절한 관리	④ 외부 DB·시스템 연계 시 적법근거 검토 ⑤ 추론 수준 및 입력데이터 최소화·추상화 ⑥ 복잡한 데이터처리의 가시성, 통제가능성 확보	⑦ 보다 엄격한 적법성·비례성 검토 ⑧ 부정확한 처리 등에 따른 권리침해 방지 ⑨ 자동화된 결정에 따른 권리침해 방지

1 기초업무 보조

○ (개요) **LLM을 활용**한 **자료 가공·요약·검색, 챗봇 상담** 등 **일상적인 업무**를 보조하는 **가장 기초적이고 일반화**된 유형

- AX 목적*을 고려할 때 **상대적으로 프라이버시 위험이 낮을** 수 있으나, **내부문서 입력, 프롬프트 수집·처리** 등에 따른 **유출·침해 위험 존재**

* 대규모 개인정보를 분석하거나 권리에 영향을 미치는 의사결정을 직접 지원하는 기능이 제한적

○ (중점 검토) **안전한 활용 환경 조성, 기본적 안전조치** 적용

- **허용·금지되는 용례** 등을 포함한 **이용지침** 안내·교육, **이용이력 점검 등 내부통제** 체계 마련
- **레드팀 테스트** 등을 통해 프롬프트 인젝션, 탈옥, 데이터 추출 공격 등 **외부공격**에 대한 **안전성 검증·개선**
- **AI 입력·출력 데이터 저장 여부, 저장 위치 및 보관기간** 명확화, 개인정보 **위수탁 및 국외이전** 여부 검토 및 **규정 준수**

※ 민간 서비스 이용 시 사업자 목적으로 데이터가 활용(AI 모델 학습 등)되지 않도록 처리위탁의 형태로 계약을 체결하고, 서버 위치에 따라 국외 이전 여부 검토

2 정보 연계 · 분석 · 추천

○ **(개요) 맞춤형 정보 분석, 추천** 등을 통해 **인간의 사고, 의사결정 지원**

- 서로 다른 맥락으로 **흩어져있던 정보**가 **하나의 시스템에서 연계**되면서 **개인정보 유출, 목적 외 이용 위험**이 **확대**될 수 있고,
- **개인의 특성, 행동, 선호** 등이 체계적으로 **분석·평가**될 수 있음

○ **(중점 검토) 목적 외 이용, 개인정보 추론 리스크** 집중 관리

- **외부 DB·시스템을 연계**할 때 **처리의 목적·대상·범위의 확대·변경**되지 않는지 **면밀히 검토**하고 **처리 근거 재검토**
- **분석·추론되는 정보의 유형, 수준**을 목적 달성에 필요한 범위로 **한정·추상화**하는 등 **사생활 침해, 개인정보 처리 최소화** 노력
- 연계된 DB·시스템은 업무 목적에 따라 **접근권한을 차등 부여**하고, **접근통제, 접속기록 모니터링** 등 **강화된 관리체계** 구현

3 선별 · 판단

○ **(개요)** AI가 **행정 수혜자 선별, 위험 탐지** 등 **공공 의사결정 과정에 직접적으로 활용·연계**되는 **선별, 판단 결과**를 제공

- **자동화된 처리**가 정보주체의 **권리·의무에 영향**을 미칠 수 있고, **과도한 감시가 우려**되는 등 프라이버시 **리스크가 상대적으로 확대**

○ **(중점 검토) 엄격한 적법성 검토, 부정확 처리 및 자동화** 리스크 관리

- AX의 **구체적 행정 목적을 정의**하고 **필요성·효과성을 사전 검증**, **개인정보처리 최소화**, **법령상 근거** 등 **명확한 적법근거** 확보 필요
- AI의 **편향, 정확성 부족**(inaccuracy), **강건성**(robustness) **미흡** 등으로 인한 **정보주체 권리침해를 예방**하기 위한 조치를 마련·이행

※ 개인정보 등 입력변수의 품질·대표성·정확성, 알고리즘 설계·성능 등 점검·개선

- **보호법상 자동화된 결정** 여부를 검토하고, 해당할 경우 **거부권, 설명·검토 요구권**의 **실질적 보장**을 위한 대응체계 등 마련

※ 시스템 구조상 해당하지 않더라도, 행정 담당자가 AI 결과를 기계적으로 수용하는 등 사실상 완전히 자동화 결정으로 운영되지 않는지 실태점검 등을 통해 확인 권장

3 추진 체계

○ (각 공공기관) **기관장·CPO·CAIO 등을 중심으로** 개인정보보호를 **AX의 핵심 요소**로 인식하고 **내부 관리체계** 정비

- CPO에게 **권한 및 역할을 부여**하고, CPO 및 개인정보보호 부서, AX 현업부서, 정보화 부서 등 **관련 부서 간 협력체계** 구축

- **AI 운영환경을 고려**하여 **내부 관리계획 고도화, 민감정보·대규모 개인정보 처리**가 수반되는 경우 **개인정보 영향평가*** 수행

* 영향평가를 수행하지 않더라도 AI 특화 영향평가 항목에 대한 자체 검토 등 고려

○ (보호위원회) **공공 AX 프라이버시 헬프데스크**를 통해 직접 지원, **가이드라인 등 AI 규범** 지속 정립, **AI 특례 도입** 등 현행법상 한계 개선

공공 AX 혁신지원 헬프데스크	**ˆ 법령 해석, ˆ 사전적정성 검토제, ˆ가명처리, ˆ규제샌드박스** 등 **적합한 지원 수단과 연계**하는 **종합창구**
AI 분야 가이드라인	**에이전틱 AI, 피지컬 AI** 등 새로운 기술환경 속 **적법근거 해석, 리스크 유형·경감방안** 등을 **지속 업데이트**
AI 특례 도입	**강화된 안전성 확보**를 전제로 **공익·사회적 이익증진 AI 기술개발 목적**의 개인정보 **원본 활용 허용** 가능하도록 **보호법 개정**

Ⅳ. 향후계획

○ 공공 AX 프라이버시 헬프데스크 상시 운영(계속)

○ 에이전틱 AI 시스템 관련 적법기준, 리스크 경감 방안 안내(연내)

○ 'AI 특례' 도입을 위한 보호법 개정 추진(계속)

⇒ **(부처 협조 사항) 각 기관별 자율적 검토 및 보호 노력을 추진하는 한편, 불확실성 등으로 지원이 필요할 경우 보호위원회 헬프데스크 적극 활용**

과학기술관계장관회의	
회 차	2026 - 7 (5호)

공공 AX
프라이버시 보호 안내서(안)

2026. 7. 23.

개인정보보호위원회

목 차

Ⅰ. 개요

1 발간 배경 및 목적

□ **인공지능 전환**(AX)이 공공부문에서 본격화되며 **국가 운영 방식**과 **행정 서비스 체계**를 **근본적으로 재편**

○ AI 도입으로 행정은 **더 빠르고 정확**해지며, 국민은 **맞춤형·선제적 공공서비스**를 제공받는 **유능하고 효율적인 정부**로 변화

※ 정부 AX 사업 예산은 '25년 0.5조원→'26년 2.4조원으로 약 5배 확대(과기정통부, '26.1.)

※ 국정과제(24번. 세계 최고 AI 민주정부 실현), AI 행동계획(전략7)에 공공AX 과제 포함

- 정부는 범정부 AI 공통기반 구축, 공공AX 프로젝트, 공공 인공지능 지원사업 등 적극 추진

□ **공공기관**은 **국민의 삶 전반**을 포괄하는 **방대한 데이터를 보유**하고 있으며 **대국민 서비스·정책 집행**을 통해 **강력한 영향력**을 행사

○ 공공기관은 **법령상 국가의 책무 이행, 공익 수호** 등을 전제로 국민의 주민등록번호, 건강·소득 정보 등 **개인의 신원과 권리, 사회적 지위**에 직접적으로 영향을 미치는 **광범위한 개인정보를 처리**

○ 민간 대체가 어려운 **필수·독점적 성격**으로 인해 국민의 서비스 **선택권이 실질적으로 제한**되고 **개인정보 수집**을 **회피하기 어려운 구조**

□ 이에, **공공 AX**의 **안전성**은 **정부 정책의 신뢰에 직결**되며, 프라이버시 보호는 **국민의 신뢰 확보, 기본권 보장, 위험 통제**의 **핵심 수단**

○ 설계 단계부터 **개인정보를 적법하고 안전하게 처리**하기 위한 **명확한 기준**과 **대응 방안**을 수립하여 **국민이 수용하는 공공 AX** 추진 필요

□ 보호위원회는 본 안내서를 통해 **공공 AX 추진 시 개인정보 처리 이슈를 체계화**하고, **법적 기준과 안전조치, 의결 사례** 등을 **함께 제시**

○ 이를 통해 **공공기관의 책임있는 AX**를 **적극 지원**하고자 함

※ 안내서는 향후 현장 의견, 적용 사례 등을 반영하여 지속 고도화할 예정임

2 적용 범위 및 성격

□ 본 안내서는 **공공기관**[1]이 **AX를 추진**하는 과정에서 **개인정보 처리가 수반**되어 **개인정보처리자의 지위**를 갖는 경우 적용되며, **기관별 상황·맥락**에 따른 **자율적 활용을 위해 마련**되었음

○ 아울러, 공공 AX 시스템의 **개발·구축** 또는 **운영 업무**를 **위탁받은 외부 업체 등 수탁자** 등도 본 안내서를 참고할 수 있음

□ 본 안내서는 **AI와 관련**된 **기존의 보호위원회 안내서**를 기반으로 하되 **공공부문 특성을 중심**으로 **보완 설명**한 것으로서,

○ 본 안내서를 일차적으로 활용하면서 **필요시 개별 안내서**의 **구체적, 세부적 내용을 확인**할 것이 권장됨

< 보호위원회 주요 관련 안내서 >

안내서 명	주요 내용
AI 개발·서비스를 위한 공개된 개인정보 처리 안내서('24.7.)	AI 개발·서비스 과정에서 공개된 개인정보 처리의 법적 근거, 안전성 확보 조치, 권리보장 방안 등 제시
AI 프라이버시 리스크 관리 모델('24.12.)	AI의 프라이버시 리스크 요인을 체계화하고 리스크 관리의 방향과 원칙을 제시
생성형 AI 개발·활용을 위한 개인정보 처리 안내서('25.8.)	생성형 AI 수명주기 각 단계에서 고려해야 할 법적 기준, 안전조치 등을 제시
자동화된 결정에 대한 정보주체의 권리 안내서('24.9)	자동화된 결정에 대한 정보주체의 권리행사(보호법 §37의2) 조치 사항, 자동화된 결정의 범위 및 기준과 절차 등의 공개에 관한 사항 등 안내
개인정보 처리 통합 안내서('25.7.)	개인정보 처리 시 준수해야할 사항을 전반적으로 안내
개인정보의 안전성 확보조치 기준 안내서('25.11.)	개인정보처리자가 조치하여야 하는 최소한의 안전성 확보기준을 안내
생성형 AI 서비스 이용자를 위한 개인정보 보호 가이드('26.5.)	생성형 AI 이용 과정에서 국민이 마주할 수 있는 8가지 주요 질의에 대한 답변, 관련 수칙 제시

□ 공공AX와 관련된 『**개인정보 보호법**』 준수 가능성을 높이기 위한 것으로서 **다른 법령상 규율에 대해서는 다루지 않음**

1) 중앙행정기관 및 그 소속 기관, 지방자치단체, 국회, 법원, 헌법재판소, 중앙선거관리위원회의 행정사무를 처리하는 기관, 그 밖의 국가기관 및 공공단체 중 대통령령(제2조)으로 정하는 기관을 주요 독자층으로 상정함(보호법 §2제6호)

참고하기

【 공공분야 AI 도입에 대한 해외 정책·가이드라인 사례 】

1 **『M-25-21: Accelerating Federal Use of AI through Innovation, Governance, and Public Trust』** (백악관 예산관리국(OMB), '25.4.)

- **(개요)** 美 연방기관의 AI 활용, 거버넌스, 공공신뢰 확보를 위한 행정지침
- **(주요내용)** 연방기관이 임무수행, 대국민 서비스에 AI를 적극 활용하도록 권고하면서 프라이버시, 시민권, 시민자유 보호 등 공공신뢰 확보를 병행하도록 요구
  - 기관별 Chief AI Officer 지정, AI 거버넌스 체계 구축, AI 활용사례 관리, 위험관리 절차 마련 등 기관차원의 책임체계 정비 촉구

2 **『Orientations for ensuring data protection compliance when using Generative AI systems (Version 2)』** (유럽 개인정보감독기관(EDPS), '25.10.)

- **(개요)** EU 기관·기구·사무소·청(EUIs)의 생성형 AI 활용시 개인정보보호 의무 이행을 지원하기 위한 실무 가이드
- **(주요내용)** 개인정보 처리흐름 식별, 목적 구체화 및 적법근거 검토, 개인정보 최소화, 고위험/대규모 처리시 영향평가(DPIA) 수행, 공급자 관리 등 안내

3 **『AI Playbook for the UK Government』** (영국 디지털서비스국(GDS), '25.2.)

- **(개요)** 중앙정부, 공공기관 등이 AI를 안전하고 책임있게 활용할 수 있도록 기본원칙, 위험관리, 보안, 개인정보보호, 인간통제 기준 안내
- **(주요내용)** AI의 가능성뿐 아니라 오류, 환각, 편향 등 한계를 이해한 활용, 보안 및 개인정보보호, 중요 업무 AI 활용 시 인간검토 보장 등 제시

4 **『Advisory Guidelines on Use of Personal Data in AI Recommendation and Decision Systems』** (싱가포르 개인정보보호위원회(PDPC), '24.3.)
※ 공공부문에 한정된 지침은 아님

- **(개요)** AI 기반 추천·예측·결정 등 인간 의사결정자를 지원하는 시스템에서 개인정보를 활용하는 경우 관련규정 적용 기준을 AI 생애주기별로 제시
- **(주요내용)** 적법근거 확보, 통지·동의 및 투명성 확보, 책임있는 활용을 위한 내부 정책·절차 마련 및 공개, AI 검증도구 및 영향평가 활용 등 안내

참고하기

【 공공부문 AI·알고리즘 도입의 프라이버시 이슈 사례 】

과거 해외에서는 **공공부문 AI·알고리즘 적용** 과정에서 **권리침해, 차별적 편향, 과도한 감시** 등 **프라이버시 논란**이 불거진 바 있다. **최근**에는 기술 고도화로 **자동화 수준 및 데이터 결합·분석 범위가 확대**되고 있으며, 블랙박스 특성으로 **판단 과정을 설명하기 어려운 문제**가 더욱 커지고 있다. 따라서 과거 **사례에서 확인된 위험을 더욱 중대하게 고려**할 필요가 있으며, **위험 수준에 비례한 강화된 통제와 책임 구조**를 마련할 필요가 있다.

1 호주 복지급여 과·오지급 선별 사례

- **(개요)** 호주 정부는 **복지급여 과·오지급 여부 확인**을 위해 개인의 **연간 소득자료와 복지급여 수급자료**를 **자동 대조**하는 시스템을 운영('16~'20)
  - **부적절한 알고리즘***으로 **무고한 시민에게 채무가 부과**되는 등 **자동화된 산정 결과**가 **사실상 불이익 처분의 근거**로 활용되면서 **집단소송, 조사**로 이어짐

    * 비정규직·시간제 노동자의 불규칙한 소득을 기계적으로 연간 평균화하여 계산 등
- **(시사점) AI·알고리즘의 결과**가 국민의 **권리·의무에 직접 영향**을 미치는 경우 **자동화된 산정 결과**를 **행정처분의 근거로 그대로 활용**하는 것은 **신중**할 필요
  - **수반되는 개인정보 처리**의 **목적 적합성, 정확성**에 대한 검증과 함께, **인간의 실질적 검토, 설명 제공, 오류 정정 절차** 등 권리구제 장치를 필수적으로 결합할 필요

2 네덜란드 복지급여 사기 예측 사례

- **(개요)** 네덜란드 지자체는 **복지급여 사기 예방·적발**을 위해 **데이터 프로파일링 기반**의 **위험 예측 시스템**(SyRI)을 도입('17~'21)
  - 직접적 사기 증거가 아닌 **성별, 가족상황, 혼인상태, 언어 능력*** 등 **인적 정보를 위험도 평가에 활용**하면서 **취약계층**에 대한 **감시, 차별 논란** 대두

    * (예) '젊은 나이', '싱글맘(한부모 가정)', '네덜란드어 구사 능력 부족' 등
- **(시사점)** 최종 처분 단계 이전의 **데이터 수집·활용 단계**에서도 **과도한 프로파일링, 감시, 차별·편향** 등 **프라이버시 침해가 초래**될 수 있음
  - **특정 집단**에 대한 **프라이버시 침해**, **불균형한 처분**이 발생하지 않도록 **시스템의 입력 변수, 학습데이터, 산출 방식** 등에 대한 **면밀한 설계 필요**

3 영국 공공장소 AI 얼굴인식 사례

- **(개요)** 영국 경찰은 공공장소에서 **실시간 얼굴인식 기술**을 사용해 **통행인의 얼굴 이미지**를 **수배 중인 인물 목록과 대조***하는 시스템 운영('16~)

  * 잠재적 일치 결과가 나오면 현장의 경찰관이 확인해 적법하게 조치

  - **시민단체**는 공공장소에서의 생체정보 처리가 **대량감시를 유발한다고 비판**
  - **영국 개인정보 감독기구(ICO)**는 경찰이 얼굴인식 기술을 사용하는 경우 **권리와 자유를 존중**하고 **데이터보호 영향평가, 보호조치, 투명성, 책임성 요건**을 갖춰야 한다고 강조

- **(시사점) 공공안전 목적**이더라도 **불특정 다수 국민을 대상**으로 한 **비자발적 개인정보 처리**가 수반될 수 있으므로 **엄격한 법적 근거와 비례성 심사 필요**

  - 알고리즘의 **편향·오탐 가능성** 검증, 시스템이 사용되는 **상황 및 추적 범주 명시**, **인간의 최종 판단 절차**, **보관·삭제 정책** 등 다각적인 통제장치 마련 필요

참고하기

【 AI 프라이버시 보호의 기본 원리 】[2)]

1 리스크 기반(risk-based) 접근

- **개별 AI의 용례와 유형**에 따라 **개별 맥락을 고려**하여 리스크를 식별, 평가해 대응하는 **맞춤형, 핀셋형 접근**을 지향(일률적 통제 지양)
- **리스크**에 대한 통제 강도는 평가된 **리스크 수준에 비례**하여 **점진적으로 상향**

2 원칙 기반(principle-based) 접근

- **빠르게 변화하는 기술 환경**을 고려하여 개별 기술에 대해 세세하게 정하기보다는 **원칙 중심으로 접근**하여 **기술 중립성** 견지

【 개인정보 보호 원칙 】 (法 제3조)

① 처리 목적 명확화 및 목적에 필요한 최소 수집
② 목적 내 적합한 처리 및 목적 외 활용 금지
③ 개인정보의 정확성·완전성·최신성 보장
④ 기술적·관리적·물리적 보호조치 등 안전관리
⑤ 처리사항 공개 및 정보주체 권리 보장
⑥ 사생활 침해 최소화 처리
⑦ 익명·가명 처리의 원칙
⑧ 책임준수 및 신뢰확보 노력

3 사전 예방 중심 체계 전환

- 데이터 경제 전환으로 **개인정보 활용 규모·범위가 확대**되면서 **사고를 온전히 막기 어렵고**, **사고 후 제재**만으로는 피해 **예방·회복에 한계**
- 이에, **위험을 사전에 식별·관리**하는 **예방형 체계**로 전환하여 **개인정보 보호가 기본**이 되는 안심사회(Privacy by Default) 구현

2) 『AI 프라이버시 리스크 관리 모델('24.12.)』, 『예방 중심 개인정보 관리체계 전환 계획('25.12.)』 참고

Ⅱ. 공공 AX 추진 단계별 검토

◈ **① 사전 설계, ② 개발·구축, ③ 적용·관리 단계별 핵심사항**을 검토하여 **AI 수명주기 전반에 걸친 프라이버시 보호** 체계를 확립합니다.

※ **(고려사항) 사업 계획수립** 단계에서 **주관 공공기관**이 全 단계의 **요구사항을 사전 검토**하고, **개발업체 등 협력 기관과 협의**함으로써 원활한 사업 추진을 도모할 수 있습니다.

【 이것만큼은 꼭 점검하세요 공공 AX 프라이버시 10대 핵심 점검 】

사전 설계 단계	1	AX·개인정보 처리를 통해 달성하고자 하는 **구체적 목적을 정의**하세요.
	2	목적 달성에 필요하고 확보 가능한 **개인정보 항목을 확인**하세요. ※ AI 모델 개발 및 시스템 구축, 시스템 운영, 시스템 고도화 등 목적별로 확인
	3	개인정보 항목별로 수집·이용·제공의 **적법근거를 확인**하세요.
	4	**AI 개발·구축 방식***을 정할 때, 성능뿐만 아니라 활용되는 **개인정보의 민감도**, 기관의 **개인정보보호 역량 및 여건 등을 고려**하세요. * 상용 서비스 연계, 공개모델 고도화, 자체 개발 / 추가학습, RAG 등
개발·구축 단계	5	**데이터 수준에서 안전조치**를 적용하세요. ※ (예) 학습데이터 및 RAG 데이터 가명·익명처리 등 데이터전처리
	6	**AI 모델 수준에서 안전조치**를 적용하세요. ※ (예) 차분 프라이버시 등 개인정보보호강화기술(PETs)을 활용한 학습 등
	7	**시스템 수준에서 안전조치**를 적용하세요. ※ (예) 개인정보 입·출력 필터링, AI 최소권한 부여 등
시스템 적용·관리 단계	8	배포 전·후 **AI 안전성 테스트**(AI red-teaming) 등을 통해 개인정보 유·노출 가능성을 지속 점검·개선하세요.
	9	**개인정보 처리방침**을 통해 개인정보 수집 사실, 처리 목적, 처리 과정 등을 투명하게 알리세요.
	10	**정보주체 권리보장** 체계를 구축하세요. * 열람, 정정·삭제, 처리정지 요청 등에 대한 신고채널 구현, 대응절차 마련 등

1 사전 설계 단계

○ **AX·개인정보 처리를 통해 달성**하고자 하는 **구체적·합리적 목적을 설정**하는 것은 프라이버시를 고려한 AX의 출발점

- AI가 활용되는 **목적, 맥락, 대상, 기대효과** 등을 명확히하여 **목적 제한, 최소수집** 등 **보호 원칙이 준수**될 수 있도록 하고(보호법 §3),

- **목적 달성에 필요**하고 **확보 가능**한 **개인정보 목록을 체계적으로 식별**하여 **수집·이용·제공** 등의 **적법근거를 확보**해야 함(보호법 제3장)

○ 개인정보는 **▲AI 모델 개발 및 시스템 구축**, **▲시스템 운영**, **▲시스템 고도화** 등 **AX 전 주기**에 걸쳐 활용될 수 있으므로 **단계별·목적별로 구분**하여 **적법근거를 확인**할 필요

<AX 단계별·목적별 개인정보 활용 예시>

AI 모델 및 시스템 개발 · 구축	√ AI 시스템을 구성하는 **AI 모델을 학습**하기 위한 데이터 (AI 모델 미세조정(fine-tuning) 등) √ AI 모델과 연계되어 **AI 시스템을 구성하는 외부 데이터 구축** 등을 위한 데이터(RAG 벡터 DB 등)
AI 시스템 운영	√ AI 시스템이 실제로 운영되기 위하여 **수집, 생성, 저장 등 처리되는 입력 및 출력 데이터**(LLM 입·출력 프롬프트, 문서, 영상 촬영 데이터, 통화 음성데이터 등) √ 사용자가 **개인화된 RAG 구축**을 위해 업로드하는 데이터
AI 시스템 고도화	√ AI 서비스 제공과정에서 **수집된 이용자 데이터**를 해당 AI **시스템의 고도화를 위해 AI 모델 학습** 등에 활용하는 경우 등

- 적법근거 검토는 처리목적·출처별로 다를 수 있으나, **수집 목적 내 활용***(보호법 §15①), **합리적으로 관련된 추가 처리**(보호법 §15③), **가명정보 처리 특례**(보호법 §28의2)** 등을 검토해 볼 수 있음

* △ 동의, △ 법률에 특별한 규정 및 법령상 의무 준수, △ 공공기관 소관 업무 수행, △ 계약 체결·이행, △ 급박한 생명·신체·재산 이익, △ 정당한 이익, △ 공공의 안전·안녕

** 개인정보를 안전하게 가명처리하면 △통계작성, △과학적 연구, △공익적 기록보존 목적에 한하여 정보주체 동의 없이 활용할 수 있도록 한 제도

- 법에서 특별히 제한하고 있는 **민감정보, 고유식별정보, 주민등록번호(보호법 §23, §24, §24의2)** 처리 시에는 **처리제한 예외에 해당**해야 함

참고하기

【 공공 AX 개인정보 처리의 적법근거 해석】

■ **공공업무 수행 과정**에 **AI를 도입**하는 경우 추가적인 개인정보 처리가 발생할 수 있으나, **모든 AX가 새로운 개인정보 처리를 수반하는 것은 아님**

- **다수의 공공 AX**는 **담당자가 직접 수행**하거나 **전통적 ICT 시스템**(규칙·키워드 기반 등)을 통해 처리하던 업무를 **AI 시스템으로 대체·고도화**하는 형태에 해당

- 이때 **개인정보 처리의 목적, 처리 대상, 정보주체 범위** 등이 **변동되지 않는다면 결과적으로 기존과 동일한 개인정보 처리**가 이루어짐

※ 다만, AI 시스템은 AI 모델, 데이터 저장소 등으로 구성된 디지털 시스템으로서 개인정보 처리의 적법성과 별개로 사이버보안 위협 등에 대응하기 위한 보다 충실한 안전성 확보 조치가 필요함을 유념

■ **기존의 처리행위**가 법령상 근거 등 **적법한 처리 근거에 기반**하고 있고, AI 도입이 기존 **처리 범위를 실질적으로 확대하지 않는다면 기술 중립적 관점**에서 **기존의 적법근거가 계속 적용**될 수 있음

- 한편, **업무 수행 전·후의 AI 시스템 개발·고도화** 과정은 구체적 맥락에 따라 **기존 적법 근거 적용 가능성**을 검토하거나, **합리적으로 관련된 추가적 이용** 및 **정당한 이익** 적용 가능성, **가명정보 활용 특례** 적용 등을 검토할 수 있을 것임

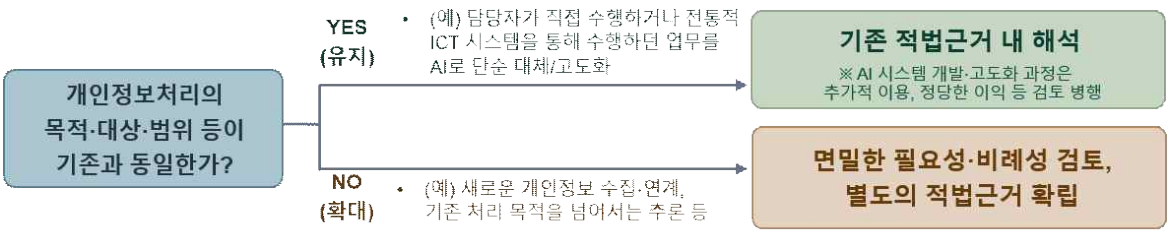
※ **(추가적 이용)** ① 수집 목적 관련성, ② 예측 가능성, ③ 이익 침해 여부, ④ 안전조치 고려
(정당한 이익) ① 목적의 정당성, ② 처리의 필요성, ③ 이익형량 고려

■ 반면, AI 도입으로 **개인정보 처리의 범위, 목적 및 영향이 확대**되는 경우* 처리의 **필요성·비례성을 면밀히 검토**하고 **별도의 적법근거를 확립**해야 함

* (예) ▲기존에는 필요하지 않았던 개인정보를 새롭게 수집하는 경우, ▲ 여러 시스템에 분산된 개인정보를 통합·분석하여 개인에 대한 특성, 성향 등 새로운 정보를 추론하는 경우, ▲ 기존 처리 목적을 넘어 개인에 대한 평가·예측·선별 결과를 생성하는 경우 등

■ 결론적으로, 공공 AX의 개인정보 처리 적법성은 **AI 활용 여부 자체가 아니라** AI 도입으로 인해 **기존 개인정보 처리의 목적·범위·영향이 실질적으로 변화하는지 여부**를 중심으로 판단할 필요가 있음

< 공공업무 AI 전환 시 적법근거 검토 방향 >



관련 사례

【 수집 목적 내 활용 해석 사례 : 범부처통합연구지원시스템(IRIS) 고도화】

※ '26.3. 보호위원회 사전적정성 검토 의결

■ **(상황)** R&D **평가위원 및 협업 연구자 추천 방식**을 기존 키워드·기술분류 중심에서 AX연구성과·신조어·융합 연구 등 **의미·맥락 중심으로 고도화**

- 이를 위해 **IRIS에 축적된 연구개발정보**를 **AI 모델 학습, RAG DB 구축** 등에 활용

■ **(근거) 법률에 특별한 규정(보호법 §15①2)**

- 『**국가연구개발혁신법**』 보호법 §17(연구개발성과의 활용), 보호법 §19(연구개발정보의 처리 등), 보호법 §20(통합정보시스템의 운영 및 이용) 및 시행령·고시(『국가연구개발정보처리기준』) 등 법률에 특별한 규정이 있는 경우로 해석

- 혁신법 보호법 §20에 의한 시스템을 고도화하여 기존 기술로는 불가능했던 최신 정보를 AI 학습하고 강화된 검색, 조회 기술을 통해 구현하는 것으로, **시스템의 구축 및 운영·이용 범위를 벗어나지 않는 한** 법상 **처리 목적 범위 내**에 포함

- 게시판 이용자 대상 **사전고지·의견수렴**, **상시적 점검**, AI 모델 지속 고도화, **처리위탁 요건 충족** 등 안전조치 부과

관련 사례

【 수집 목적 내 활용 해석 사례 : 온라인 게시글 기반 수도사고 사전인지】

※ '26.4. 보호위원회 사전적정성 검토 의결

■ **(상황) 제휴 맘카페 게시물**을 수집하고 **상용 LLM(API 연계)으로 내용을 분석**해 단수, 녹물, 적수, 벌레 유충 등 **수도사고 위험 징후***를 **조기에 탐지**하여 **초기 대응**(피해 범위 추정, 원인 규명 등)에 활용

* (예) "적선동 물이 안나오네요" → 단수 탐지, "냄새나는 물이 나와요" → 녹물 탐지

- 수도 관련 글이 게시될 가능성이 있는 **특정 게시판 내 게시글만 수집**하고, 내용에 포함된 식별성 높은 **개인정보는 마스킹**

■ **(근거) 처리자의 정당한 이익(보호법 §15①6)**

- **안정적 생활용수 공급**이라는 **합법적·구체적 목적의 정당성**, 수도사업 특성상 **기존 방식의 한계** 등을 고려한 **수단의 합리성**, **권리침해 최소화 조치*** 등을 고려해 처리자의 정당한 이익으로 해석

* 게시판 이용자 대상 사전고지 및 의견수렴 절차 이행, 상시적 점검, AI 모델 지속 고도화, 처리위탁 요건 충족 등 안전조치 부과

관련 사례

【 추가적 이용 해석 사례 : 이용자 대화 데이터의 AI 모델 학습 】

※ '25.5. 보호위원회 사전적정성 검토 의결

■ **(상황) LLM 성능 개선**을 위해 **이용자 프롬프트 입력** 내용을 **AI 학습데이터로 수집·이용**

■ **(근거) 합리적으로 관련된 범위 내 추가적 이용(보호법 §15③)**

- LLM의 환각(hallucination), 편향(bias) 등 **리스크를 완화하고 성능을 높이기 위해** 이용자와의 상호작용 데이터가 **학습에 활용될 필요**가 있으며 이는 LLM **서비스 운영과 밀접하게 관련**된 점,
- LLM 서비스는 이용자의 **질문 맥락을 파악하여 통계적으로 적절하다고 선정된 답변 문구를 생성**하는 그 **성질상 기존 대화 내용이 학습데이터로 활용될 수 있음을 예측가능**하고, 학습데이터 **수집 사실 및 거부 방법**(opt-out)을 **대화창 알림으로 수차례 고지**하여 **예측 가능성**을 높이는 점,
- 대화 데이터는 모델 **학습 데이터로만 활용**되고 구축되는 **통계모형 자체**로는 정보주체에게 **직접적인 영향을 미치지 않으며**, 이용자가 **학습데이터 수집을 항시 거부**(opt-out)할 수 있는 기능도 제공하여 정보주체의 **이익을 부당하게 침해하지 않는 점**,
- 학습 데이터셋 내 **개인식별 가능성이 높은 정보를 탐지·삭제**하는 **필터링** 절차를 운영하는 점 등을 고려하여 **추가적 이용**으로 해석

관련 사례

【 추가적 이용 해석 사례 : 보이스피싱 의심번호 DB 구축·활용】

※ '25.7. 보호위원회 사전적정성 검토 의결

■ **(상황) 전화 수발신 내역 데이터**를 활용하여 **보이스피싱 의심번호를 예측**하고 이를 **금융사의 이상거래 탐지·차단**에 이용

- 보이스피싱 **범죄자의 통화·문자 수발신 패턴을 학습**한 AI 모델을 생성

■ **(근거) 합리적으로 관련된 범위 내 추가적 이용(보호법 §15③)**

- 보이스피싱은 통신망을 악용한 금융사기로, **이용자 보호 의무 준수**를 위해 이를 **사전에 예방하고 차단**하는 것은 **정상적인 통신·금융 서비스 제공과 밀접하게 관련**된 점,
- 대다수 통신사·금융사는 **기존에도** 스팸·이상거래 방지 등 **부정이용 및 이용자 보호 목적의 서비스를 운영 중**이며, **이용약관 등을 통해 이를 안내**하고 있어 **고객보호 의무의 연장선**으로 보이스피싱 예방·탐지 서비스에 대한 **예측이 가능한 점,**
- 통시사가 예측한 **의심번호의 정·오탐 여부**를 **금융사가 별도 검토하고 피드백**하는 절차가 구축되어 있어 **정보주체의 이익을 부당하게 침해하지 않는 점,**
- 보이스피싱 의심 DB 구축 및 통신사, 금융사 간 통신 시 **전화번호가 암호화되어 저장 및 송·수신**되는 점 등을 고려해 추가적 이용을 근거로 해석

관련 사례

【 추가적 이용 해석 사례 : 활동 데이터 기반 맞춤형 검색결과 제공】

※ '26.5. 보호위원회 사전적정성 검토 의결

■ **(상황)** 이용자의 **검색 서비스 이용기록**에 더하여, **블로그·카페 글에 대한 활동기록***, **쇼핑 이력** 등 사업자가 제공하는 인접 서비스의 이용내역 데이터를 **AI로 분석**하여 **개인화된 답변 생성**에 활용

* 게시글 등에 대한 좋아요·공유 등 활동기록으로, 전체공개된 콘텐츠에 한함(부분공개·비공개 등 일반 공중의 접근·열람이 허용되지 않는 콘텐츠는 수집 범위에서 제외)

■ **(근거) 합리적으로 관련된 범위 내 추가적 이용(보호법 §15③)**

- **종합 포털 서비스의 성격** 및 **기존의 검색 결과 제공 방식** 등을 고려한 **수집 목적 관련성**, 처리방침·이용약관 및 검색결과 화면 구성 등을 통한 **예측 가능성 확보,** 거부 옵션 제공 등 데이터 활용에 대한 **통제권 부여 및 안전성 확보 조치**를 고려하여 추가적 이용으로 해석

- 다만, **거부 옵션의 명확한 안내,** 맞춤 검색 **서비스 주요내용의 투명한 공개**, 맞춤정보 **추상화 처리 및 주기적 삭제,** 맞춤정보 DB에 대한 **강화된 접근권한 관리·접근 통제** 등 **추가적 안전조치 부과**

관련 사례

【 민감정보 처리 제한 예외 해석 사례 : 119구급 현장지원 시스템 고도화 】

※ '26.7. 보호위원회 사전적정성 검토 의결

■ **(상황) 구급대원의 수기 작성, 판단에 의존**하는 **▲병원 전 응급환자 중증도 분류**(Pre-KTAS), **▲적정 이송병원 판단**, **▲구급일지 작성**을 AI로 **자동화·지능화**

- 이를 위해, **환자의 건강 정보 등 민감정보가 포함**된 **구급대원, 환자 및 보호자간 대화 내용**(텍스트 변환 데이터)을 전처리하여 **AI 시스템에 입력**

■ **(근거) 법령상 민감정보 처리 요구·허용(보호법 §13①2)**

- 『**119법**』 **§22**(구조·구급활동의 기록관리), 『**응급의료법**』 **§31조의4**(환자의 중증도 분류 및 감염병 의심환자 등의 선별) 및 하위 규정에서 **처리되는 민감정보의 종류를 열거***하고 **처리를 요구·허용**하고 있는 경우로서,

* 법정 서식에 민감정보 기재 사항이 있는 경우도 포함

- AI를 통해 **기존 업무 수행 방식**을 **기술적으로 고도화**하는 과정에서 **처리 목적, 처리 범위**가 **변동되지 않으므로** 기존과 동일하게 법령상 처리가 요구·허용되는 범위 내에 해당

※ AI 관련 데이터별 저장·이용·보관·파기 등 관리기준 설정, 시스템별 접근권한 설정 및 접근통제, AI 모델 성능관리 등 안전조치 부과

○ AX의 목적을 설정하고 개인정보 처리의 적법 근거를 식별했다면 **상용 서비스 연계, 자체 개발** 등 **AI 개발·구축 방식을 구체화**하게 됨

- 이때, AI 성능뿐만 아니라 **활용되는 개인정보의 민감도, 기관의 ICT·보안 역량** 등을 종합적으로 고려하여 **프라이버시 보호와 AI 활용 효과를 균형**있게 확보할 수 있음

- 방식에 따라 **데이터에 대한 통제 수준, 개인정보 처리위탁** 및 **제3자 제공** 여부, **보안·운영 부담** 등이 상이하므로 사전에 검토

< AI 개발·구축 방식별 주요 검토사항 >[3]

상용 서비스 연계	공개모델 고도화	자체 모델 개발
■ 개인정보 처리위탁 또는 제3자 제공 해당 여부 - 접근통제, 암호화, 파기 등 보호조치를 확인하고, - 서비스 제공기관의 학습활용 배제 등을 계약으로서 보장하는 Enterprise API 우선 고려	■ 모델의 라이선스·이용조건을 확인하고, 보안 취약점 등 주요 업데이트 지속 반영 ■ 추가학습 데이터에 포함되는 개인정보 최소화, 적법성 확보	■ 모델 학습, 운영 전 과정의 안전성 확보, 지속적 유지관리 ■ 대규모 학습데이터의 적법성 확보, 개인정보 최소화 등 AI 학습 전반에 대한 통제

※ 외부 기관에 대한 개발·구축 업무 위탁 시에도 개인정보 처리위탁 여부 확인, 규정준수 필요

○ **업무 특화**를 위해 **검색 증강 생성**(RAG[4])을 활용할 때에는 검색 대상 데이터의 **개인정보 최소화** 및 **정확성·최신성** 확보, **접근권한 설정, 적법근거 확보** 등을 종합적으로 고려

참고하기

【 검색증강 생성(RAG) – 추가학습(fine-tuning) 프라이버시 리스크 분석 】

■ **(검색증강 생성)** 참조되는 **벡터 DB에 포함된 개인정보**가 **검색·조합 과정**에서 **출력에 그대로 유·노출**될 위험 등이 존재함

- 다만, 개인정보 **처리가 해당 데이터 참조 시점으로 한정**되고, 열람·삭제 등 정보주체 **권리보장이 상대적으로 용이**

■ **(추가 학습) 추가학습 데이터에 포함된 개인정보**가 **모델 파라미터에 암기**되어 **출력에 재현**될 수 있다는 분석이 있으나,

- **학습을 위한 토큰화, 임베딩** 등 전처리 과정에서 **개인에 대한 식별성이 낮아질 수 있고**, 생성형 AI 대비 **판별형 AI**에서는 **관련 위험이 낮아질 수 있음**

3) ▲ 상용 서비스 연계 : 비공개 모델 등 상용 AI 서비스를 활용(예: API)하는 방법으로 신속한 개발 가능
▲ 공개모델 고도화 : 주로 공개(open-weight)된 사전학습 모델(pre-trained model)을 활용해 AI 시스템을 개발하는 방식,
▲ 자체 모델 개발 : 모델을 처음부터 직접 학습하는 방식으로 고비용 인프라 및 전문 인력 등 자원이 요구됨

4) Retrieval-Augmented Generation : AI의 성능 향상을 위해 외부 지식베이스를 검색하고 이를 입력값에 결합하는 기법

2 개발 · 구축 단계

○ 설정한 목표·전략에 따라 **AI 시스템을 실질적으로 개발·구축**하면서 **프라이버시 리스크 완화**를 위한 **안전조치**를 **내재화**해야 함(보호법 §29)

○ AI는 전통적인 개인정보 침해, 유·노출 위협에 더하여 **AI 모델 학습, 사용자 입력 수집, 추론 및 RAG 연계 등 시스템 동작**에 따른 AI 특유의 개인정보 처리 지점별 **새로운 리스크를 유발**

- 이에, 시행령(보호법 §30) 및 관련 고시*에서 규정하고 있는 안전조치 외에도 **리스크 유형 및 경감방안에 대한 논의 동향**을 지속적으로 살피며 **데이터·모델·시스템 수준**에서 **안전조치를 보완**해 나갈 필요

* 『개인정보의 안전성 확보조치 기준』: 접근권한, 접근통제, 암호화, 접속기록의 보관 및 점검, 악성프로그램 등 방지, 물리적 안전조치, 내부 관리계획 수립 등 규율

< AI의 주요 프라이버시 리스크 및 경감방안 예시 >[5]

개인정보 처리 지점	주요 프라이버시 리스크	리스크 경감방안 예시
대규모 학습 데이터 구축·활용	불필요한 개인정보·민감정보 학습, 목적 외 활용 등	데이터 최소화 원칙 준수, 가명·익명처리, 차분 프라이버시 기법[6] 적용 등
입력 프롬프트 수집	불필요한 개인정보 입력, 내부 문서 복사 입력 등	입력 제한·필터링, 자동 마스킹, 금지되는 입력 안내, 사용자 교육 등
RAG·지식베이스 연계	개인정보 포함 문서의 벡터화, 권한 없는 데이터 접근, 개인정보 검색 결과 노출 등	문서 등급 분류, 민감 문서 제외, 권한 기반 검색, 검색 로그 감사 등
모델 출력 생성	학습데이터 암기 및 재현, 악의적 프로파일링 등 개인정보 추론 등	출력 필터링, 출력 제한 및 모니터링, 호출 횟수 제한 등
로그·이용기록 저장 및 처리	장기 보관에 따른 프로파일링, 고도화 목적 재사용 등	보존기간 제한, 목적별 분리, 학습 재사용 제한, 접근권한 통제 등
외부 API/도구 연계, 업무 위·수탁	수탁자 자체 모델 학습 등 통제 부족, 목적 외 처리 등	학습활용 제한 등 위수탁 계약 조항 검토·체결, 최소 권한 부여 등
에이전트·외부도구 호출	권한 초과, 데이터 전송 통제 부족, 목적 외 이용 등	최소 권한 부여, 사람 승인 절차 마련, 로그기록 관리, 자동실행 제한 등

5) 『AI 프라이버시 리스크 관리 모델('24.12.)』, 『생성형 인공지능 개발·활용을 위한 개인정보 처리 안내서('25.8.)』 및 『OWASP Top 10 For LLM Applications('24.11.)』 등 국내외 자료를 참고하여 재구성

6) 차분 프라이버시(Diffential Privacy;DP) : 데이터 세트에 노이즈를 추가하여 분석 결과를 도출함으로써, 해당 분석 결과로부터 특정 개인을 역으로 추론하지 못하도록 하는 기술. AI 학습·파인튜닝 과정에 차분 프라이버시를 적용하여 개별 학습 샘플이 노출되지 않도록 노이즈를 관리.

참고하기

【 에이전틱 AI와 프라이버시 침해 위험 】

■ **에이전틱 AI**는 복잡한 **다중 작업**을 **자율적으로 수행**하기 위해 **LLM 위에 구축**될 수 있는 **자율 시스템**으로, 최근 **AI의 중요한 발전 양상임**

- AI 에이전트 역시 **기존 LLM의 데이터 흐름 단계를 포함**하지만 **높은 자율성, 외부 도구·시스템 연계, 장기 기억** 등으로 인해 **추가적인 복잡성**을 초래

■ 이에 따라, **국내외 사이버보안·개인정보보호 관련 기관**에서 AI 에이전트와 관련된 **새로운 위협과 경감방안**에 대한 **논의가 진행**되고 있음

- **현재 공공 AX 부문**에서 **에이전틱 AI 시스템**을 구현한 사례는 **제한적**이나, 향후 발전에 대비하여 살펴볼 수 있음

< AI 에이전트의 취약점 및 프라이버시 리스크 예시 >[7)]

취약점 유형	개요	프라이버시 리스크	리스크 경감 방안
에이전트 목표 해킹 (Agent Goal Hijack)	프롬프트 조작, 숨겨진 명령 등을 통해 에이전트의 목표, 의사결정 경로 변조	**의도하지 않은 개인정보 처리·유출**	프롬프트 주입 방지, 최소권한 원칙 적용, 중요 작업시 인간 승인 요구 등
도구 오용 및 악용 (Tool Misuse and Exploitation)	모호한 지시, 불안전한 위임으로 도구를 잘못 사용	**데이터 유출, 과도한 정보 접근**	도구별 기능 및 데이터 접근 범위 제한, 도구 호출 시 명시적 인증 요구 등
신원 및 권한 남용 (Identity and Privilege Abuse)	저장된(cached) 자격증명, API 키, 검색 데이터 유출·재사용	**권한 오남용 및 무단 접근**	작업 범위·시간 제한, 에이전트 정체성과 컨텍스트 분리 등
메모리 및 컨텍스트 독성 (Memory&Context Poisoning)	공격자가 메모리, 컨텍스트를 오염시켜 안전하지 않은 행동 유도	**데이터 유출 및 잘못된 도구 사용**	전송 데이터 암호화, 최소권한 접근, 콘텐츠 검증 등
인간·에이전트 신뢰 악용 (Human-Agent Trust Exploitation)	사용자와 형성된 신뢰관계를 악용해 민감정보 추출, 기만행위 수행	**사회공학적 개인정보 탈취**	로그 보관 및 감사체계 구축, 의심 상호작용 신고채널 운영 등

7) 『OWASP Top 10 For Agentic Applications 2026('25.12.)』 일부 내용을 발췌하여 재구성

3 시스템 적용 · 관리 단계

○ 공공 AX 시스템을 **실제 적용·관리**하면서 **지속적으로 프라이버시 리스크를 점검·개선**하고 **정보주체의 권리를 보장**해야 함

○ 공공 AX 시스템을 **배포·적용하기 전** 프라이버시 **리스크가 적절히 식별·경감**되었는지 등 **안전성을 최종 점검**

- **실환경에서의 테스트**(red-teaming 등)를 통해 **외부 공격 저항성, 학습 데이터 유·노출 가능성** 등을 점검하고,

- AI 모델, 소스코드, 학습 데이터, 외부 도구 연계(API 등) 등 **시스템 전반에 대한 쟁점을 최종 검토·분석**

○ AI 시스템은 **배포 이후**에도 **사용자의 이용행태, 새로운 공격 기법 등장, 모델 업데이트, 외부 API 정책 변경** 등으로 위험 양상이 달라질 수 있으므로 **지속 점검 및 유지관리 필요**

- **주기적 테스트**, 산출물의 개인정보 포함 여부 등 **출력 샘플링 감사**, 프롬프트 인젝션 등 **이상 행위 탐지, 로그 보존·모니터링 체계** 등 구축

○ **투명성 확보**를 위해 **데이터셋 수집 사실, 주요 출처, 처리 목적, 개인정보 처리 과정** 등을 **개인정보 처리방침**(보호법 §30), **FAQ** 등에 공개

참고하기

【 생성형 AI 서비스의 개인정보 처리방침 작성 시 고려사항 】

※『개인정보 처리방침 작성지침('26.4.)』 본문 및 부록(생성형 인공지능(AI) 서비스 처리방침) 참고

■ 생성형 AI 서비스는 복잡한 개인정보 처리과정을 수반하므로 정보주체가 쉽게 이해할 수 있도록 알기쉬운 용어로 작성한 별도의 처리방침을 제공하는 것을 권장함

■ 또한 AI의 특성을 고려해 다음 사항이 분명히 드러나도록 기재하는 것이 바람직

√ 해당 AI 서비스의 의도된 용례(intended use)
√ 정보주체가 입력한 정보 및 생성된 결과물의 수집 여부 및 보유·이용기간
√ 수집된 데이터의 해당 AI 모델·시스템 고도화 학습활용 여부
√ 정보주체의 학습 거부권 행사방법 및 절차

○ 개인정보 **열람, 정정·삭제, 처리정지**(보호법 §35, §36, §37) 정보주체의 **권리를 실질적으로 보장**할 수 있는 **기술적·관리적 조치를 마련·실행**

- 개인정보 처리와 관련한 **신고·문의 및 의견제출 창구**를 제공하고, 권리침해 **모니터링 및 대응절차**를 마련

- **대응절차**를 마련할 때는 AI 시스템의 구조적 특성으로 인해 일부 제약이 있을 수 있으나, **시간·비용·기술적 측면**에서 **합리적으로 실현 가능한 범위**에서 최대한 보장

참고하기

【 실현 가능한 범위에서의 정보주체 권리보장 노력 예시 】

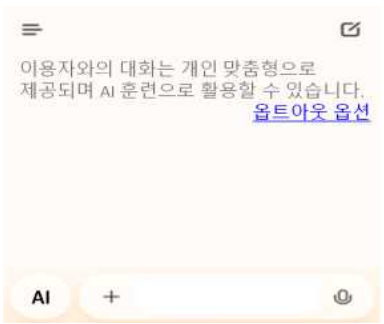
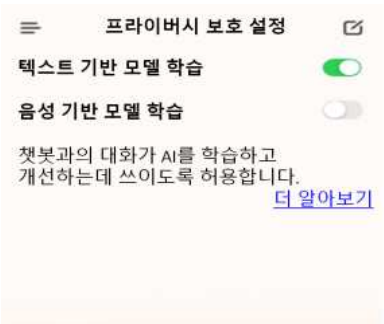
■ AI는 정형화된 데이터베이스와는 달리 토큰화·임베딩 등 전처리[8], 모델 파라미터를 통한 지식 표현 등 기술적 특성으로 인해 개인정보의 위치를 특정하거나 개별 정보만을 식별·삭제하는 데 현실적 어려움이 존재할 수 있고, 이로 인해 정보주체 권리보장이 일부 제약될 수 있음

- 그러나 AI 개발자 및 서비스 제공자는 당시의 기술 수준(State of the art), 시스템 구조 및 개발 방식 등을 고려하여 정보주체의 권리행사를 지원하기 위한 정책적·기술적 조치를 적극적으로 마련할 필요

■ (예시) LLM 학습에 대한 개인정보 처리정지·삭제 요청

- AI 학습 목적의 이용자 대화 이력 처리 정지 요청이 가능하도록 정보주체가 쉽게 접근 가능한 방식으로 요청 방안(Opt-out 등)을 마련하고 안내

- 학습 데이터셋의 크기, 구성 방식·체계 등을 감안하여 출력 필터링을 우선 적용하고 추후 학습 데이터셋에서 개인정보 제외

학습된다는 사실과 옵트아웃을 알기 쉽게 프롬프트 입력란에 안내 (**충분한 기간 최초 고지**, 이후 **정기적 재고지**)	**옵트아웃**을 쉽게 수행할 수 있도록 **직관적인 설정 인터페이스**를 구성
	

8) 텍스트 토큰화 : 텍스트를 컴퓨터가 이해할 수 있는 숫자 벡터(vector)로 변환하기 위하여 최소 단위(토큰)으로 분할.
토큰 인덱싱 : 각 토큰을 수치 벡터로 변환하여 토큰 인덱스를 생성. 간의 의미 유사성 등 관계를 포착하기 위한 과정
토큰 임베딩 : 각 토큰 간의 의미 유사성 등 관계를 포착하기 위한 과정

Ⅲ. 공공 AX 유형별 검토

◈ AX 단계별 검토사항에 더하여 **AX 유형별 차등화된 점검을 통해 위험 수준에 기반(risk-based)한 맞춤형 보호체계로 고도화**합니다.

※ **(고려사항)** 하나의 공공 AX 시스템이 **여러 유형의 특성을 함께** 가질 수 있으므로 기관별 **AX 목적과 처리 맥락**을 고려해 관련 **점검사항을 조합하여 적용하세요.**

→ (예) '선별·판단→정보 연계·분석·추천→기초업무 보조' 순으로 **보다 엄격한 점검사항을 기준**으로 하되, 해당하는 다른 유형의 점검사항도 함께 확인하세요.

【 유형별로 집중 점검하세요 공공 AX 유형별 맞춤형 점검 】

기초업무 보조	**"안전한 활용 환경 조성, 기본적 안전조치에 중점"**	
	1	**이용지침** 안내 등 **안전한 이용 환경을 구축**하세요. ※ 이용지침 마련·교육, 이용 이력 및 접속기록 모니터링 등 내부통제
	2	**오남용, 외부 공격** 등에 대한 시스템 **안전성을 확보**하세요. ※ 악의적 프롬프트 차단 및 출력 필터링, 레드팀 테스트 등
	3	**AI 입·출력** 데이터를 **적절히 관리**하세요. ※ 저장 여부·위치·기간 설정 및 관리, 처리위탁 및 국외이전 규정 준수 등
정보 연계·분석 ·추천	**"목적 외 이용 및 개인정보 추론 리스크 집중 관리"**	
	4	**외부 DB·시스템 연계** 시 **적법근거**를 검토하세요.
	5	**분석·추론**의 **수준 및 입력 데이터**를 필요한 범위로 **최소화·추상화**하는 등 과도한 프로파일링 및 사생활 침해를 방지하세요.
	6	복잡화된 데이터 처리 과정의 **가시성, 통제 가능성** 확보 ※ AI 처리 사실 공개, 연계된 DB·시스템에 대한 접근권한·접근통제 등
선별·판단	**"엄격한 적법성 검토, 자동화된 결정에 따른 리스크 관리"**	
	7	**명확한 법령상 근거 규정** 등 **AI 기반 선별·판단**의 **적법성, 비례성을 보다 엄격히 검토**하세요.
	8	**부정확한 처리** 등에 따른 정보주체 **권리침해를 방지**하세요. ※ 개인정보 등 입력변수의 품질·대표성·정확성, 알고리즘 설계·성능 점검 및 개선
	9	**자동화된 결정**에 따른 정보주체 **권리침해를 방지**하세요. ※ 거부권·설명·검토 요구권 보장, 자동화 편향 방지, 입력변수·알고리즘 검증 등

□ 공공 AX의 유형 분류

○ **공공 AX의 유형**은 **활용 분야, 정책 목적** 등 **다양한 기준**에 따라 분류될 수 있으며*, **기술 발전**에 따라 **유연하게 변동**될 수 있음

* (예) ▲대민서비스(민원처리, 정보제공, 재난안전, 보건복지), 업무효율화 (SPRi, '25.4.), ▲환경일반, 고용노동, 사회복지일반, 공공질서 및 안전, 일반행정, 법제 (행안부, '26.3.)

○ 다만, **개인정보보호 관점**에서는 **AI 시스템**이 개인정보를 **어떠한 목적으로 처리**하고 인간의 **의사결정 과정**에서 **어떠한 역할을 수행**하는지 등 **개인정보 처리의 특성**에 따라 분류함으로써,

- 보호법상 **원칙 적용**, 정보주체 **권리침해 영향판단** 및 **권리보장, 안전조치 방안** 등에 대해 **유의미한 검토**를 수행할 수 있을 것임

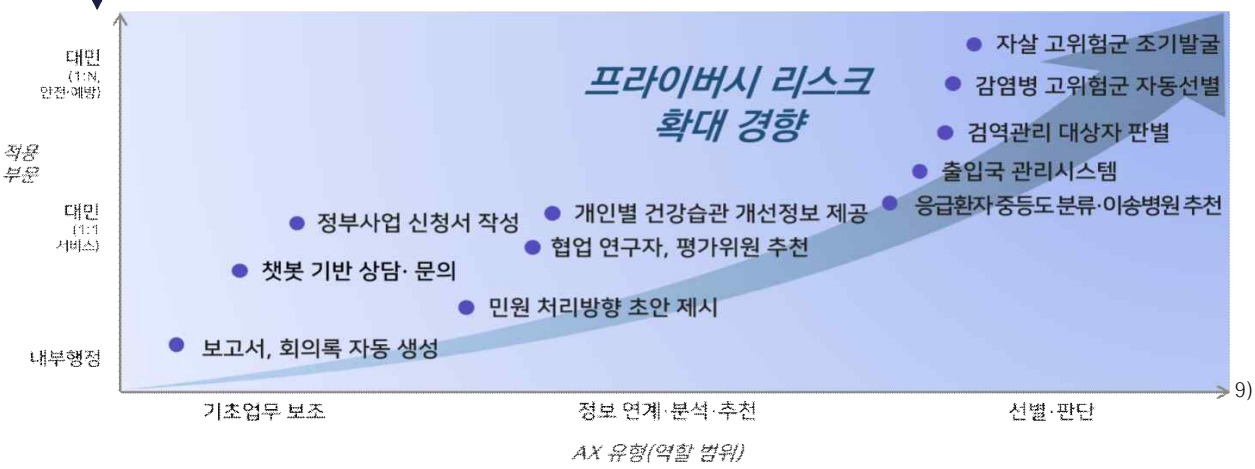
○ 이에 본 안내서는 **AI의 역할 범위, 개인정보 처리 특성** 등을 기준으로 공공 AX를 다음의 **세 가지 유형**으로 구분함

※ 하나의 시스템이 여러 유형에 중복해당될 수 있으므로 AX별 맥락에 따라 유연한 해석 필요

< 공공 AX 유형 분류 및 예시(안) >

정보주체 영향 ↓ 상대적 확대 경향

유형	AI의 역할	개인정보 처리 특성
기초업무 보조	검색, 요약, 문서 작성 보조 등 **일상적 업무 수행** 지원	기존 개인정보 처리의 효율화
정보 연계 분석·추천	다양한 데이터 연계에 기반한 **분석·예측·추천** 결과 제공	새로운 데이터 결합 및 추론 정보 생성
선별·판단	대상자 선별, 위험 탐지 등 **직접적인 의사결정** 지원	처리 결과가 구체적 행정조치와 연계



9) 1:N은 불특정 다수를 대상으로 하는 선제적 개입형(감시, 예방 등), 1:1은 개별 민원인 대상 응답형 행정 업무를 상정함. 프라이버시 리스크는 AX 유형(x축)과 적용부문(y축)의 조합으로 달라지며, 특히 두 축이 동시에 상승할 때 프라이버시 리스크가 확대되는 경향이 있음

1 기초업무 보조

□ 개요

○ **LLM을 활용**한 **자료 가공·요약·검색, 챗봇 상담** 등 **일상적인 업무**를 보조하는 **가장 기초적이고 일반화**된 유형

- 복합적인 AI 시스템도 기초업무 보조 기능을 일반적으로 포함하므로 **최소한의 개인정보 보호 기준으로 참고**할 수 있음

○ **이미 상당 부분 일반화**된 사례로, **필요한 활용은 적극 지원**하면서 **안전한 이용 환경을 조성**하는 것이 중요함

- 특히 공공부문은 **책임성·신뢰성이 요구**되는 만큼 기초 수준이더라도 **기관의 공식적인 관리체계** 아래 **안전하게 전개**하는 노력이 필요함

※ 적절한 관리 체계가 부재할 경우 소속기관의 승인·통제 없이 개인이 임의로 이용하는 비인가 AI 활용에 따른 우려 존재

□ 프라이버시 리스크 분석

○ **대규모 개인정보를 분석**, 권리에 영향을 미치는 **의사결정을 직접 지원**하는 기능이 **제한적**이므로 프라이버시 **위험이 상대적으로 낮을 수 있음**

○ 다만, **개인정보가 포함된 내부 문서, 프롬프트** 등을 **입력·처리**하는 과정에서 **개인정보 유출, 목적 외 이용** 등의 위험이 발생할 수 있음

- **내부 활용** 시 **적절한 이용지침, 접근권한 관리 및 통제가 미흡**할 경우 **불필요한 개인정보 입력, 민감정보 처리, 무단 접근 등** 우려

- **대국민 시스템**의 경우 **불특정 다수가 이용**하는 환경으로 인해 **학습 데이터 및 RAG 데이터 추출** 등 **외부 공격**에 노출될 **가능성이 확대됨**

○ **정보주체 권익 영향이 상대적으로 제한적**이므로 업무 특성, 처리하는 정보의 성격을 고려해 **상용 AI 서비스 활용**도 **유연하게 검토**할 수 있음

- 이 경우, 개인정보의 외부 전송·처리가 수반될 수 있고, **처리위탁, 제3자 제공 및 국외이전** 해당 여부와 필요한 보호조치를 충분히 검토하지 않으면 개인정보 **무단 제공·이용 위험**이 발생할 수 있음

□ 중점 검토

“안전한 활용 환경 조성, 기본적 안전조치 적용”

① 이용지침 안내 등 안전한 이용 환경 구축

○ **허용·금지되는 AI 용례, 개인정보 처리** 기준 등을 포함하는 **이용지침**을 마련하여 시스템 주요 **사용자를 대상으로 안내·교육**

○ **대화창 등 사용자 인터페이스**를 통해 불필요한 **개인정보·민감정보 입력 주의** 등 AI 활용 **유의사항을 알기 쉽게 안내**

○ 내부 활용 시 직원의 시스템 **이용이력, 접속기록 등을 점검**하는 등 시스템 **오남용 예방**을 위한 **내부통제 체계** 마련

② 시스템 오남용, 외부 공격 등에 대한 시스템 안전성 확보

○ 챗봇 상담 등 대민 AI 서비스는 **프롬프트 인젝션, 학습데이터 추론·추출 공격**[10] 등 **외부공격에 대한 안전성**을 보다 강화할 필요

- **입력 필터**를 통해 **악의적인 프롬프트를 탐지·차단**하고, **출력 필터**를 통해 **개인정보, 민감정보**의 **부적절한 출력 방지**

- **AI 레드티밍**[11] 등을 통해 **배포 전·후 안전성**을 검증하고, 발견된 취약점은 **모델 개선, 입·출력 필터링 고도화, 안전 정렬**(safety alignment) 등을 통해 개선

○ **AI 챗봇**이 계정 정보 조회·변경 등 **확대된 권한을 독자 수행**하지 않도록 **권한을 적절히 제한**하고, **중요한 행정처리**는 **사람의 확인 절차 유지**

※ (관련 이슈) 공격자가 메타 AI 지원 챗봇을 속여 새 이메일을 등록하고 비밀번호를 재설정하여 인스타그램 계정을 탈취('26.6월)

---

10) 프롬프트 인젝션(prompt injection) : 악의적인 지시가 포함된 사용자 입력이나 외부 데이터를 통해 AI가 기존 지침과 다른 동작을 수행하도록 유도하는 공격.
학습데이터 추론·추출 공격 : 특정 입력이나 반복·조작된 질의를 통해 특정 데이터가 학습데이터에 포함되어 있는지를 유추하거나 학습데이터를 재구성하려는 악의적 시도.

11) AI 레드티밍(AI Red-teaming) : AI 시스템이 어떤 위험을 일으킬 수 있는지 사전에 점검하는 일종의 ‘모의 공격 테스트’로서, 사람 또는 다른 AI가 AI 모델을 의도적으로 시험해 보며 유해 콘텐츠 생성, 편향된 답변, 보안 우회(jailbreak) 등 문제를 식별하는 도구·활동. 외부 전문인력 레드티밍, 자동화 레드티밍 등 다양한 방식 존재.
구체적 방법론은 과기정통부·KISA 『AI 보안 레드티밍 가이드('26.7.)』 등을 참고할 수 있음

참고하기

【 AI 레드티밍 사례 – 금융보안원 금융분야 점검 】[12)]

■ **'25년 주요 점검대상**

① **금융소비자 대상 챗봇 서비스** : 상담, FAQ 응답, 금융 정보 안내 등
② **임직원 업무 지원 서비스** : 뉴스/공시 요약, 정보 검색, 문서 분석 등
③ **금융 연계 서비스** : 자연어 기반 금융 서비스

■ **평가 방법론**

(1) 적대적 공격 방어 관리체계 점검

데이터 오염 공격 대응 체계	- 데이터 무결성 확보(위변조 방지, 이상치 탐지, 데이터 정제) - 신뢰할 수 있는 출처에서 데이터 수집
모델 오염 공격 대응 체계	- 모델 무결성 확보(위변조 방지, 형상 관리, 검증 평가) - 사전학습 모델 출처 확인 및 안전한 파일 포맷 사용
데이터·모델 추출 공격 대응 체계	- 부적절하게 수집된 개인정보 포함 여부 확인 - 입출력 횟수 제한 및 민감정보 노출 방지
회피 공격 대응 체계	- 적대적 공격 데이터 활용 사전학습 및 주기적 테스트 - 방어 대책(추가학습, 필터링, 가드레일) 구현 - 외부 입력값 실시간 탐지 및 대응 체계

(2) 벤치마크 및 자동화 도구 기반 실증 점검

※ (벤치마크) 외부 논문 및 커뮤니티를 통해 수집한 벤치마크, 자체 생성 벤치마크 등
(자동화 도구) PyRIT 등 사용 가능하나 망분리 환경, 낮은 시간 효율성 등으로 제한적

(3) 수작업 레드티밍 실증 점검

공격 시나리오		데이터·모델 정보 추출*, 유해 콘텐츠 생성 유도, 허위 정보 및 조작 생성 유도, 시스템 기능 훼손, 금융서비스 기능 제어 우회 * 학습·참조데이터 추출, 개인정보 및 민감정보 유출 시도 등
공격 기법	**직접 명령 주입**	지시 내용 무시 유도, 역할극 기반 우회 기법
	맥락 조작	가상의 시나리오 및 상황 설정, 긴급 상황이나 특수 상황 가정, 교육적/연구적 목적으로 포장한 요청
	입력 난독화	문자 인코딩 변경, 토큰 문자 및 특수문자 삽입, 희귀 언어 사용, 은유적 표현이나 암시적 언어 사용
	다단계/연쇄 공격	여러 단계에 걸쳐 점진적 경계 확장, 이전 응답을 근거로 한 추가 요청, 여러 프롬프트의 조합을 통한 우회

12) 금융보안원『2025 FSI AI RED TEAM REPORT('25.12.)』

③ 적절한 AI 입·출력 데이터 관리

○ **AI 입력·출력 데이터** 등 시스템에서 처리되는 데이터의 **저장 여부, 저장 위치 및 보관기간**을 명확히 관리하고 안전하게 파기

○ **위수탁 계약 체결, 처리방침을 통한 공개** 등 개인정보 **처리위탁**(보호법 §26) 및 **국외 이전**(보호법 §28의8) 관련 **법적 의무 준수** 필요성 검토

- **민간 AI 서비스**를 이용하는 경우 **입력 데이터**가 **사업자의 AI 모델 학습 등에 활용되지 않도록** 이용 약관, 계약 조건 등 확인

* AI 학습활용 배제 규정을 둔 **기업용 API(Enterprise API) 활용** 우선 고려

관련 사례

【 서비스형 LLM을 활용한 진료대화 처리('24.10. 사전적정성 검토 의결)】

■ 의료기관이 **진료 대화를 기반**으로 **의료기록 작성업무를 자동화**하는 과정에서 **개인용 무료 라이선스**로 서비스형 LLM을 사용하면 **입력 데이터**가 LLM 서비스 **제공자의 자체 목적(예: LLM 학습)으로 활용될 우려**가 있음

⇒ **기업용 라이선스(Enterprise API)**와 같이 데이터를 **해당 의료기관의 목적으로만 처리**하고 LLM 서비스 제공자의 목적으로 처리하지 않는 방식으로 이용하도록 함

2 정보 연계 · 분석 · 추천

□ 개요

○ AI가 기초업무 수준을 넘어 **맞춤형 정보 분석, 추천** 등을 통해 **인간의 사고, 의사결정을 지원**하는 수준

○ **다양한 정형·비정형 데이터가 수집·활용**되거나, **개별 업무 시스템에서 관리**되던 데이터가 **하나의 AI 시스템**에서 연계·분석될 수 있음

- 여러 데이터가 결합·분석됨에 따라 개별 데이터만으로는 **파악하기 어려웠던 관계, 패턴, 특성** 등 **새로운 정보**가 도출됨

- **LLM, 머신러닝 등 AI 기술의 발전**으로 **문맥 기반**의 분석, 추론이 가능해지면서 규칙·키워드 기반 방식에 비하여 **분석의 범위가 크게 향상**되고 **보다 정교한 행정 지원이 가능해짐**

□ 프라이버시 리스크 분석

○ **개별 업무에서 제한적으로 활용**되던 개인정보가 **새로운 분석이나 서비스에 활용**되는 과정에서 **목적 외 이용** 리스크 존재

- 특히 **복지, 조세, 보건, 민원** 등 **서로 다른 행정영역**의 정보를 결합하는 경우, **개별 법령상 허용된 범위** 등 **기존의 처리근거를 넘어** 활용될 수 있으므로 면밀한 검토 필요

- 또한, 다양한 **데이터베이스, 외부 시스템이 연계**되는 과정에서 개인정보 **처리 흐름이 복잡화**되고, **투명성 저하 및 보안취약점** 및 설정 오류가 개인정보 **침해로 이어질 가능성** 증가

○ AI를 통해 **명시적으로 드러나지 않은 개인의 특성, 행동, 선호** 등이 **정보주체가 합리적으로 예상**하기 어려운 방식으로 **추론될** 위험

- 자신의 개인정보가 **어떠한 방식으로 연계·분석**되고, **어떠한 추론정보가 생성·활용**되는지 충분히 **인지하거나 통제하기 어려울 경우 개인정보 자기결정권 약화**로 이어질 수 있음

□ 중점 검토

“목적 외 이용, 개인정보 추론 리스크 집중 관리”

① 데이터 연계·분석의 목적 적합성 및 적법근거 검토

○ **외부 DB·시스템을 연계**할 때는 개인정보 **처리의 목적·대상·범위가 확대 및 변경**되지 않는지 **면밀히 검토하고,**

- 확대·변경 시 목적 외 이용이 없도록 **추가적인 처리 근거**를 확보

② 과도한 프로파일링 및 사생활 침해 방지

○ 분석·추론에 사용되는 **입력 데이터**는 **반드시 필요한 개인정보에 한정**하는 등 처리되는 **개인정보 최소화**

○ **분석·추론의 수준**은 업무 목적 달성에 **필요한 범위로 한정·추상화**하여 **과도한 프로파일링 및 사생활 침해를 방지**하고,

- 맞춤형 분석 기능 등에 대한 **거부 방법과 의미**를 **알기 쉽게 안내**

③ 복잡화된 데이터 처리 과정의 가시성, 통제 가능성 확보

○ 정보주체가 **AI 및 데이터 연계에 대해 인지**할 수 있도록 **AI 활용 사실, 추천 결과 생성 과정*** 등을 **처리방침, 사용자 인터페이스** 등을 통해 제공

* 활용 데이터의 범주, 주요 고려요인 등

○ **학습 데이터, 연계 데이터**(RAG 등)의 **정확성·최신성**을 확보하기 위해 **주기적인 데이터 갱신, 품질관리** 등 실시

○ 연계된 DB·시스템은 업무 목적에 따라 **접근권한을 차등 부여**하고, **접근통제, 접속기록 모니터링** 등 강화된 관리체계 구현

관련 사례

【활동데이터 기반 맞춤형 검색결과 제공 사례('26.5. 사전적정성 검토 의결)】

√ 이용자별 **맞춤 정보 DB**에는 **민감정보가 포함되지 않도록** 조치

√ **추론데이터 길이를 한정**하고 **주기적으로 갱신**하여 불필요한 정보 저장 방지

√ 불필요한 상세정보 추론·데이터 누적, 민감정보·고유식별정보 처리 방지를 위해 **지속적으로 모니터링**

√ 맞춤정보는 **최상위 관심사 카테고리((예시: '여행', 운동(야구)' 등)로 추상화**

√ 맞춤정보 DB에 대한 **접근권한 관리·접근통제 시 강화된 안전조치** 강구

관련 사례

【 평가위원·협업연구자 추천 사례('26.3. 사전적정성 검토 의결)】

√ **RAG 구조**를 통해 추천 시점에 **최신의 연구성과 및 이력정보를 직접 참조**함으로써 개인정보의 **정확성·최신성 확보**

√ **특정 속성**이 추천 결과에 **과도한 영향을 미치지 않도록 가중치 조정** 등 편향 통제장치를 적용하여 **차별·침해 위험 최소화**

3 선별 · 판단

□ 개요

○ AI가 정보 연계·분석·추천 수준을 넘어 **공공 의사결정 과정에 직접적으로 활용·연계**되는 **선별·판단 결과**를 제공

- **AI 도입**으로 **수집·분석**되는 **데이터 종류·양이 확대**되고, 관찰된 데이터간 **상관관계가 효과적으로 분석**되면서 **행정 수혜자 선별, 위험 탐지** 등 사회적 필요성이 높은 서비스가 가능해짐

○ 이러한 기능은 **행정 사각지대 해소, 위험의 조기 탐지, 의사결정의 일관성 보장** 등에 사회적 편익에 기여할 수 있으나, **정보주체의 권리·의무**에 **직접적인 영향**을 미칠 가능성이 **상대적으로 높음**

※ "공공서비스"가 사람의 생명, 신체의 안전 및 기본권의 보호에 중대한 영향을 미치거나 위험을 초래할 우려가 있을 경우 『인공지능기본법』상 고영향 AI에 해당

(『인공지능기본법』상 "공공서비스" 영역 정의) 공공서비스 제공에 필요한 **자격 확인, 결정 또는 비용징수** 등 **국민에게 영향**을 미치는 국가, 지방자치단체, 공공기관 등의 의사결정

□ 프라이버시 리스크

○ **AI 시스템의 결과**가 **행정처분의 근거로 활용**되는 과정에서 **부정확한 개인정보 처리** 등으로 인해 **국민에게 실질적 불이익**이 발생

- **해외**에서도 **복지급여 과·오지급 선별** 등 AI·알고리즘 기반 선별 결과의 국민 **권익 침해 논란**이 제기된 바 있음

- AI의 출력은 **데이터 기반 통계적 추론** 등에 기반한 것으로, **인간의 실질적 검토 없이 기계적으로 수용**하는 것은 **신중 필요**

○ 지속적 **모니터링**은 **개인의 행동과 활동**에 대한 **광범위한 개인정보 처리**를 수반할 수 있으며 **과도한 감시**로 이어질 우려

- 정보주체가 개인정보 **처리 사실, 범위** 등을 **합리적으로 예측하기 어려운 방식**으로 **사생활이 포함된 정보**(SNS 등), **영상정보** 등이 **광범위하게 분석**될 경우 **침해 우려** 증가

□ 중점 검토

“엄격한 적법성 검토, 부정확 처리 및 자동화에 따른 침해 방지”

① AI 기반 선별·판단의 적법근거·비례성 엄격히 검토

○ AI 기술을 통해 **효과적인 데이터 연계·분석**이 가능하다는 **이유만으로 광범위하고 과도한 개인정보 활용**이 수반되지 않도록 **유의**

- **공익적 목적이더라도 명확한 적법근거**에 기반하여 처리되어야 하며,
- AI 시스템 도입 및 개인정보처리가 **해결하고자 하는 행정 목적 달성**에 **필요**한지, 활용되는 개인정보의 범위와 처리 방식이 **목적에 비추어 과도하지 않은지** 신중하게 검토하고,

※ 예 : CCTV, 웹크롤링 등을 통한 감시 시스템은 관련성 있는 영역에 초점을 맞춰야 하며, 더 광범위한 모니터링이 기술적으로 가능하더라도 불필요한 경우 그 범위를 제한

- **동일한 목적을 달성**할 수 있는 경우에는 정보주체 **권익에 미치는 영향이 작은 대안적 방안**을 **함께 고려**할 필요

② 부정확한 처리 등에 따른 정보주체 권리침해 방지

○ AI의 **편향, 정확성 부족**(inaccuracy), **강건성**(robustness) **미흡** 등으로 인한 **정보주체 권리침해를 예방**하기 위한 조치를 마련·이행

- AI 시스템의 목적을 고려하여 **편향, 부정확성, 오탐·미탐** 등 **핵심 위험요인을 식별**하고, **개인정보 등 입력변수의 품질·대표성·정확성, 알고리즘 설계 및 성능**을 지속적으로 **점검·개선**하는 등 위험을 완화

참고하기

【판별 AI(discriminative AI)의 용례(tasks)별 핵심 이슈 식별 예시】[13]

- **(allocative AI)** 자원 기회 명예 등 **한정된 재화를 개인에게 할당**하는 데 활용되는 시스템은 **공정성**(민감도 격차 등), **설명 가능성**에 대한 면밀한 검토 필요
- **(punitive AI)** 개인에게 **제재 등 부정적인 결과를 가하는 데 활용**되는 시스템은 **부정확성**(낮은 정확성 등), **설명 가능성**에 대한 면밀한 검토 필요
- **(cognitive AI)** 컴퓨터 비전, 이미징·진단 등 **인간의 인지를 대체, 보완**하는 시스템은 **정확성, 견고성**(robustness), **QoS 격차**(과소대표 등)이 핵심 쟁점

13) Park, S. Bridging the Global Divide in AI Regulation: A proposal for contextual, coherent, and commensurable framework. Washington International Law Journal, 33(2), 2023b

③ 자동화된 결정에 따른 정보주체 권리침해 방지

○ **보호법상 자동화된 결정**(보호법 §37의2)에 해당하는지 여부를 검토하고, **해당할 경우 거부권, 설명·검토 요구권** 등 정보주체 권리를 **실질적으로 보장**할 수 있도록 **권리행사 창구, 대응체계** 등 마련

- 자동화된 결정 해당 여부는 **형식적 요건**뿐 아니라 **실질적 인적 개입**이 있는지 여부를 종합 고려하여 판단

- 시스템 구조상 해당하지 않더라도, **행정 담당자**가 **종합적 검토 없이** AI 결과를 **그대로 수용**하는 등 **실질적으로 완전히 자동화 결정으로 운영**되지 않는지 **실태점검** 등을 통해 확인하는 것이 바람직함

참고하기

※ 상세내용은『자동화된 결정에 대한 정보주체의 권리 안내서('24.9.)』참고

【보호법상 '자동화된 결정' 판단 시 고려사항】

다음 사항에 모두 해당할 경우 보호법상 자동화된 결정(보호법 §37의2)에 해당

① '**완전히 자동화된 시스템**'에 의할 것
※ 해당 결정이 인적 개입 없이 완전히 자동화된 시스템으로 이루어져야 함

② '**개인정보를 처리**'하여 이루어질 것

③ 개인정보처리자에 의해 **'정보주체의 권리 또는 의무에 영향'**을 미치는 **'결정'**일 것

④ 정보주체에 대한 **'최종적인 결정'**일 것

⑤ 해당 시스템에 의한 '**개인정보의 처리**'와 **'결정'** 사이에 **'실질적인 관련성'**이 있을 것

⑥ **다른 법률**에 자동화된 결정 관련 **특별한 규정이 없을 것**
※ 행정기본법 제20조의 '자동적 처분', 신용정보법 제36조의2의 '자동화평가' 제외

【"인적 개입"이 있는지 판단 시 고려사항】

① 형식 요건 : "정당한 권한이 있는 사람의 개입이 있을 것"

- 업무 권한이 없는 사람, 해당 업무를 수행할 능력이 없는 사람이 형식적으로 개입한다고 하더라도 실질적 인적 개입이 있다고 볼 수 없음

② 실질 요건 : "사람의 개입이 결정의 결과에 실질적인 영향을 미칠 수 있을 것"

- 사람의 개입이 결정의 결과에 실질적 영향을 미칠 수 있어야 하고,
※ 예: 사람이 시스템 결과를 단순히 기계적으로 적용한다면 인적 개입이 있다고 보기 어려움

- 개입하는 사람이 관련 정보에 접근가능하며, 개입은 최종 결정 전 관련 정보에 대한 종합적 고려가 있는 상태(적정한 시간 부여 필요)에서 이루어져야 함

관련 사례

【 온라인 게시글 기반 수도사고 사전인지 사례('26.4. 사전적정성 검토 의결)】

√ **수도 관련 글이 게시될 가능성이 있는 게시판**을 미리 **협의·선정**(화이트리스트/블랙리스트)하여 **해당 게시판 내 게시글만 수집**

√ 수도사고 **관련성이 있는 게시글만 자동 선별**하여 **모니터링**하고, 수도사고 **관련 없음으로 분류**된 게시물은 **즉시 삭제**

- 관련 없는 게시물이 저장되지 않고 즉시 삭제되고 있는지 **주기적·상시적으로 점검**하고, 수도사고 **관련글 여부를 분류**하는 **AI 모델 지속 고도화**

√ 수집한 게시글 내용에 포함된 **식별성 높은 개인정보는 마스킹** 처리

√ 정보주체가 **크롤링 사실 및 목적**을 알 수 있도록 온라인 **카페 회원을 대상**으로 **사전고지·의견수렴 절차 이행**

관련 사례

【 보이스피싱 의심번호 DB 구축·활용 사례('25.7. 사전적정성 검토 의결)】

√ **'보이스피싱 의심번호 DB'** 구축하는 과정에서 직접적인 **통화 내역은 열람하지 않고 통화패턴의 통계적 유사도만 산출**

√ 금융사가 회신받은 **위험도 수치**는 보이스피싱 탐지의 **보조도구로만 사용**하고 **실제 보이스피싱 여부는 금융사가 재확인**하여 부당한 침해 방지

√ 통신사가 산출한 **예측값에 대한 정·오탐 여부**를 **금융사가 재확인**하여 이를 **피드백**하고, **이를 기반으로 예측모델 정확도 고도화** 절차 구축

Ⅳ. 기관별 역할 및 협력체계

□ AX를 추진하는 각 공공기관

○ **기관장·CPO·CAIO** 등을 중심으로 **AX 추진**할 때 **개인정보보호**를 **핵심 요소**로 인식하고 **CPO의 권한·역할, 협력체계** 등 **내부 관리체계** 정비

- **CPO**가 **AX의 전략 수립, 개발, 운영·관리** 등 전 과정에서 **개인정보 처리의 적법성, 안전성** 확보와 관련된 **관리·감독 책임을 수행**할 수 있도록 **권한 및 역할 부여**

- **CPO 및 개인정보보호 부서, AX 현업부서, 정보화 부서** 등 관련 담당자 간 **협력체계**를 구축하여 **CPO**가 AX에 대한 **충분한 정보를 확보**하고 **적시 피드백을 제공**할 수 있도록 체계 구축

- **AI 도입**으로 인해 **추가되는 운영환경**(AI 모델 서버, DB 등)에 대한 시스템별 **접근권한 제한 및 접근통제, 접속기록 보관·관리** 등 안전조치를 포함하여 **내부 관리계획 고도화**

○ **민감정보 또는 고유식별정보, 대규모 개인정보 처리**가 수반되는 경우 **개인정보 영향평가**(보호법 §33) 수행 고려

- AI 시스템·개인정보파일이 개인정보 보호법 **시행령 제35조 기준*** **중 어느 하나에 해당**할 경우 **의무 수행 필요**

> * △ 5만명 이상의 정보주체에 관한 민감정보 또는 고유식별정보 처리 수반
> △ 내부 또는 외부의 다른 개인정보파일과 연계하려는 경우로서 연계 결과 50만명 이상의 정보주체에 관한 개인정보가 포함되는 경우
> △ 100만명 이상의 정보주체에 대한 개인정보파일
> △ 영향평가를 받은 후에 개인정보검색체계 등 개인정보파일의 운용체계를 변경하려는 경우 그 개인정보 파일(변경된 부분에 한정)

- 영향평가를 **수행하지 않더라도** AI 전환에 따른 영향이 평가될 수 있도록 **AI 특화 영향평가 항목*에 대한 자체 검토** 등 고려

* (해당지표)『개인정보 영향평가 수행안내서』- 5. 특정 IT기술 활용 시 개인정보 보호 - 5.7. 인공지능(AI) 10개 지표

□ 개인정보보호위원회

○ **직접적인 협력**을 지원하는 **'공공 AX 프라이버시 헬프데스크'** 운영('26.1.~)

- 신청 건별 **서비스 구조, 개인정보 처리 흐름, 프라이버시 리스크 분석** 등을 토대로 **▲법령 해석, ▲사전적정성 검토제, ▲가명처리, ▲규제샌드박스** 등 **적합한 지원 수단과 연계**하는 **종합창구**

< 공공 AX 프라이버시 헬프데스크 >

◈ (상담·접수) 인공지능프라이버시팀 02-2100-3072 / 02-2100-3169

구분	지원 내용
법령 해석	■ 개인정보 보호법상 **일반적 쟁점 사항** 검토·안내
사전적정성 검토제	■ **일부 불확실성이 존재**할 경우, 처리자·보호위원회가 AI 학습 방식 및 개인정보 처리 흐름을 함께 검토하고 적정한 안전조치를 부가하여 AI 혁신 지원 ☞ **보호위원회 전체회의 의결**
샌드박스 (규제특례)	■ 범죄예방, 재난, 사회적 약자보호 등 공익적 필요가 높으나, **현행법상 개인정보 처리 근거가 없거나 금지된 사항**이 수반되는 경우, 강화된 안전조치 전제로 특례 부여 ☞ **ICT 융합·산업융합 등 분야별 규제 샌드박스 제도 활용**
가명처리 지원	■ **가명처리, 가명정보 결합, 적법여부 확인, 교육·훈련** 등 지원 ☞ **가명정보 활용 지원센터, 비조치 의견서 제도 등 보호위원회 인프라 및 제도를 통해 지원**

○ **가이드라인 등 유연한 연성규범**을 통해 현장의 **불확실성 적시 해소**

- **에이전틱 AI** 등 새로운 기술환경 속 **적법근거 해석, 리스크 유형 및 경감방안** 등을 **지속 업데이트**하여 **현장에 안내**

※ ('26 하반기(안)) 에이전틱 AI 환경에서의 데이터 처리 기준을 안내하는 정책자료 발표, 에이전틱·피지컬 AI 환경을 고려한 『AI 프라이버시 리스크 관리 모델('24.12.)』 개정 등

○ **현행법상 한계의 근본적 개선**을 위해 **'AI 특례 도입'** 등 **법개정 추진**

- **AI 특례**를 통해 **공익·사회적 이익 증진**을 위해 필요한 경우 강화된 **안전성 확보**를 전제로 **AI 기술 개발** 목적의 **개인정보 원본 활용 허용**

Ⅴ. 향후 계획

○ 공공 AX 프라이버시 헬프데스크 상시 운영(계속)

○ 에이전틱 AI 시스템 관련 적법기준, 리스크 경감 방안 안내(연내)

○ ’AI 특례‘ 도입을 위한 보호법 개정 추진(계속)

개인정보보호위원회 개인정보정책국 인공지능프라이버시팀	
담당자	임수연 사무관
연락처	전 화 : 02-2100-3078 E-mail : waterkite7@korea.kr