인공지능 지침 도입에 대한 국내외 사례 비교

진보네트워크센터 미루 정책활동가

연구의 배경 및 목적

- 세계 각국은 이미 인공지능과 관련한 구체적인 제언을 내놓고 있는 상황임
- 국내에서도 인공지능 윤리를 둘러싼 학계의 논의가 이어지고 있음
- 2019.12.17. <인공지능 국가전략>을 발표하며 '사람 중심의 AI 구현'을 표방하며 국가 경쟁력을 높이기 위한 100대 실행 과제를 제시했음
- 하지만 인공지능 도입 시 발생할 수 있는 문제 등에 대한 논의는 부족한 상황임

인공지능 도입 시 발생할 수 있는 문제

데이비드 케이(David Kaye), 의사 표현의 자유 증진 및 보호를 위한 UN 특별 보고관

"인공지능은 중립적인 기술이 아니며 인권을 위협할 수 있다"(2018)

"인공지능 활용이 마이크로 타게팅과 관련 프로파일링과 개인정보 개량 수집을 조장하고 이 때 정보주체가 개인의 정보를 거의 통제할 수 없게 됨으로써 프라이버시에 문제가 생길 것"

인공지능 도입 시 발생할 수 있는 문제

캐나다 OPC (Office of the Privacy Commissioner)

- 인공지능이 예측 및 결정을 내리는 등의 개인 행동에 영향을 줄 수 있음
- AI 활용에 있어 비합법적인 편견이나 차별, 프라이버시에 대한 위협을 가져올 수 있음

인공지능 도입 시 발생할 수 있는 문제

1. 데이터에 내재한 사회적 편견으로 인해 차별이 심화되고 강화되는 문제
2. 개인의 자율성 침해
3. 대량의 개인정보·행동정보 수집으로 인한 감시 문제
4. 개인정보 재식별의 문제
5. Nudging
   ex) Cambridge Analytica 사건
6. 투명성 보장의 어려움

국내 공공기관의 신기술 도입 현황

- 2017년 행정안전부 '지능형 정부 기본 계획'
  : "스스로 일하는 정부"를 표방하며 가치 달성을 위해 '새로운 플랫폼을 구축하고 AI활용을 적극 검토'하겠다고 밝힘
- 행정안전부 '8가지 첨단 공공서비스 분류'
  1) 챗봇
  2) 빅데이터
  3) 드론
  4) 사물인터넷
  5) 혼합현실
  6) 지능형 보안관제
  7) 블록체인
  8) 클라우딩 컴퓨팅

국내 공공기관의 신기술 도입 현황

- 한국정보화진흥원 - 인공지능 면접 실시
- 서울시 민생사법 경찰단 - 사이버 범죄 대응 등을 위해 인공지능을 도입하고자 했으나 SNS검열 등의 우려로 금지되었음
- 행정 서비스의 자동화는 물론 정책 결정에도 신기술이 영향을 미칠 것으로 예측할 수 있음
  - 지능형 정부 기본계획에 따르면 **'사전에 해결하는 정부'**의 목표 중 하나로 '인공지능이 기존 정책 데이터 분석을 통해 '숨은 정책 수요'를 발굴하여 제안'하는 시스템 조성을 내세우고 있음.

국내 인공지능 관련 원칙

- 방송통신위원회 '이용자 중심의 지능정보사회를 위한 원칙'

사람 중심의 서비스 제공 원칙	서비스의 제공과 이용은 사람을 중심으로, 인간의 존엄성을 보호할 수 있는 방향으로 진행되어야 한다.
투명성과 설명가능성 원칙	서비스 체계와 작동방식이 이용자에게 중대한 영향을 미칠 경우, **기업의 정당한 이익을 침해하지 않는 범위**에서 이용자가 이해할 수 있도록 정보를 작성해야 한다.
책임성 원칙	관련 법령과 계약을 준수하고 지능정보사회 구성원들은 지속적인 의견 교환에 참여한다.
안정성 원칙	안전하고 신뢰 가능한 지능정보서비스의 개발 및 이용, 사전에 피해 복구 방안 확보 노력
차별 금지 원칙	기술 사용에 있어 사회적 다양성 고려, 알고리즘 설계, 데이터 수집 입력 및 모든 단계에 거쳐 차별적 요소의 최소화
참여 원칙	구성원들은 공적 이용자 정책과정에 차별없이 참여할 수 있음
프라이버시와 데이터 거버넌스	서비스의 개발, 공급 및 이용의 전 과정에서 프라이버시에 미치는 영향을 최소화 할 수 있도록 하고 지속적인 의견 교환에 참여

국내 인공지능 관련 원칙

- 한국정보화진흥원, '공공기관 신뢰가능 AI구현 실용가이드: OECD 권고안의 적용'

1. 포용성장, 지속가능 발전, 복지증진

   - 공공성의 확인, 사회적 차별요소 배제

2. 인간중심 공정성

   - 인간중심 가치와 공정성 촉진, 인간중심 가치 내재화

3. 투명성, 설명가능성

   - AI 시스템에 대한 투명한 정보공개, AI시스템 결과에 대한 설명

4. 보안 및 안전성

   - AI시스템의 추적 가능성 보장, 체계적인 위험접근 관리

5. 책임성

   - AI 시스템 원칙의 실현(라이프사이클에서 발생한 의사결정과 행동 문서화)

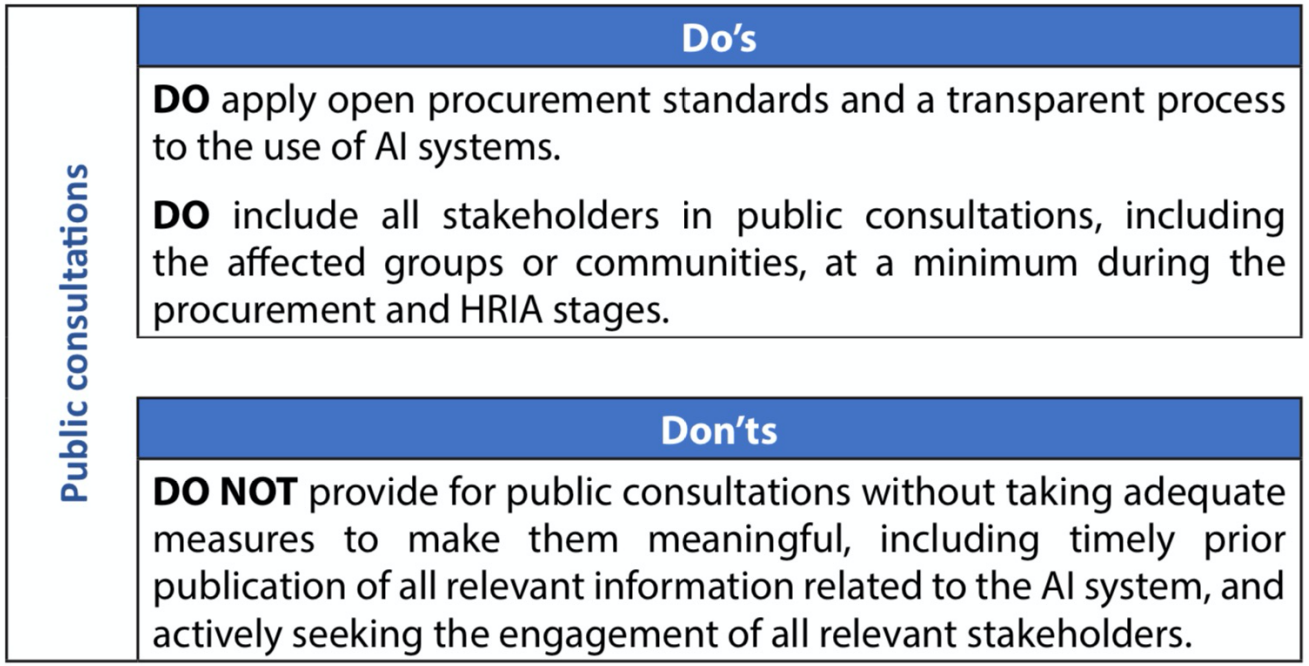
국외 인공지능 관련 원칙

- 유럽 평의회 인권위원장, ‘AI 블랙박스 해체: 인권보호를 위한 10단계

1. 인권영향평가
2. 공적협의
3. 민간부문의 인권기준 이행을 촉진하기 위한 회원국의 의무
4. 정보와 투명성
5. 독립적인 감독
6. 반차별과 평등
7. 개인정보보호와 프라이버시
8. 표현의 자유, 집회 및 결사의 자유, 노동권
9. 구제방법
10. AI 문해력의 증진

국외 인공지능 관련 원칙

- 유럽 평의회 인권위원장, ‘AI 블랙박스 해체: 인권보호를 위한 10단계는 ‘Checklists’를 포함하여 10단계를 구체화 하기 위해 ' 해야할 일’과 ‘하지 말아야 할 일’을 구체적으로 밝히고 있다.



Public consultations	Do’s
	DO apply open procurement standards and a transparent process to the use of AI systems. **DO** include all stakeholders in public consultations, including the affected groups or communities, at a minimum during the procurement and HRIA stages.
	Don’ts
	DO NOT provide for public consultations without taking adequate measures to make them meaningful, including timely prior publication of all relevant information related to the AI system, and actively seeking the engagement of all relevant stakeholders.

국외 인공지능 관련 원칙

- 유럽연합 집행위원회 인공지능 고위 전문가 그룹, ‘신뢰할 수 있는 AI를 위한 윤리 지침’

신뢰할만한 AI 구성을 위한 필수적인 요소
(AI에) 적용할 수 있는 법과 규제
윤리 원칙과 가치를 고수할 수 있는 윤리적 AI
AI 시스템이 의도하지 않은 위협으로부터, 기술적 사회적으로 모두 견고하게 버틸 수 있도록 할 것

국외 인공지능 관련 원칙

- 유럽연합 집행위원회 인공지능 고위 전문가 그룹, '신뢰할 수 있는 AI를 위한 윤리 지침'

신뢰할만한 AI를 위한 4 가지 윤리 원칙	
인간의 자율성 존중	AI 시스템 작동 전반에 걸친 인간의 개입이 있어야 함
위험으로부터의 방어	육체적·정신적 측면에서 인간의 존엄성을 보호할 수 있어야 함
공정함	AI 개발·배치·이용은 공정해야만 함: 실질적인 공정함 + 과정적 공정함
설명 가능성	과정의 투명성, AI시스템의 능력 및 목적이 공개적으로 논의될 수 있어야 함. 직간접적으로 영향을 미치는 결정들에 대해 설명할 수 있어야 함

신뢰할 수 있는 AI를 위한 요구사항 7가지	
인간 행위자의 감독	이용자에게 법적 효과 혹은 중대한 영향을 미칠 수 있는 결정에 대해 오직 자동화된 의사결정의 대상이 되지 않을 수 있어야 함
기술적 견고함과 보안	위험을 예방할 수 있는 예방적 접근, 공격으로부터의 회복력과 보안, 대비책 및 일상적 보안, 정확성, 신뢰성 및 재현 가능성
프라이버시 및 데이터 거버넌스	이용자들이 개인정보 수집 과정을 신뢰할 수 있어야 함. 데이터의 질과 온전성, 데이터 접근성 확보
투명성	추적 가능성, 설명가능성, 소통(사람들은 AI 시스템과 상호작용할 권리를 가짐)
다양성, 반차별, 공정성	불공정한 편견을 피하기, 접근 가능성 및 보편적인 설계, 이해관계자들의 참여
사회적·환경적 복지	지속 가능성 및 환경 친화적 AI, 사회적 영향 및 사회와 민주주의를 고려하는 AI
책무성(Accountability)	감사 가능성, 부정적인 영향의 최소화와 보고 체계, 각 요소들의 균형, 부정적 영향의 시정

국외 인공지능 관련 원칙

- 영국 인공지능사무국, '인공지능의 윤리와 안전을 고려한 시스템 설계·구현 가이드'
  - 영국 정부는 윤리적 가치체계를 반영한 <공공지능 인공지능 활용 가이드>에서 공공기관이 인공지능을 활용할 때 6가지 요소를 반드시 고려하도록 하고 있음

데이터의 품질
공정성
책무성
개인정보 보호
설명가능성 및 투명성
비용

결론

- 국외의 경우 이미 구체적인 법안의 형태로 제안되었지만, 한국은 여전히 원칙을 제시하는 수준에 머물러 있음.
- 인공지능이 불러올 수 있는 위험성 및 부작용을 인식하고 이에 대응하는 해외와 달리 아직 위험성 평가가 이뤄지기 전임에도 행정영역에서 이를 도입하려는 움직임이 있음.
- 인공지능의 결정으로 인한 피해 및 차별의 경우 이를 구제하거나 시정하기가 상대적으로 어렵기 때문에 이를 보완할 수 있는 도구가 개발될 필요가 있다.
- 도입 단계부터 이에 대한 고민이 필요하나, 한국에선 아직 이런 움직임을 찾을 수 없어 아쉬움이 남음