사람이 중심이 되는 인공지능을 위한

신뢰할 수 있는 인공지능 실현 전략(안)

2021. 5. 13.

관계부처 합동

요 약

Ⅰ 추진 배경

□ 인공지능이 全분야에서 빠르게 도입·활용되며 혁신을 창출하고 있으나, 확산에 따른 예상치 못한 사회적 이슈*·우려도 대두

* 인공지능 챗봇 '이루다'('21.1.), 오바마 前대통령 딥페이크('18.7.), MIT개발 사이코패스 인공지능('18.6.) 등

□ 사람 중심의 '인공지능 강국' 실현을 위해 인공지능의 혜택은 극대화하면서 위험·부작용을 최소화하기 위한 인공지능 신뢰(Trust) 확보 추진

Ⅱ 인공지능 신뢰성 개념 및 국내·외 동향

1 개념 및 주요 요소

□ **[개념]** 인공지능 윤리 실천과 이용자 인공지능 수용성 향상을 위한 핵심가치

○ 인공지능이 내포한 위험과 기술적 한계*를 해결하고, 활용·확산 과정에서의 위험·부작용을 방지하기 위한 가치 기준(요구사항)

* 귀납적 작동원리에 따른 불투명성, 학습에 활용한 데이터에 따른 편향성 등

□ **[주요 요소]** 일반적으로 안전(Safety), 설명가능(Explainability), 투명(Transparency), 견고(Robustness), 공정(Fairness) 등을 구성요소로 포함

☞ 인공지능의 개발·활용·확산에서 신뢰성 구현 기준과 방안을 정립·추진함으로써 인공지능에 대한 국민의 수용성을 향상

2 국내·외 동향

□ **[국외]** EU, 미국 등은 인공지능 신뢰성을 인공지능 윤리 실천의 핵심요소로서 강조하고 제도, 윤리, 기술 측면에서 확보방안 강구

○ **[EU]** 세계 최초로 「인공지능 법안」을 제안('21.4.)하며, 고위험 인공지능 중심의 규제(공급자 의무 부과, 적합성평가·인증 등) 제도를 선도

- 사업자의 '자동화된 의사결정 활용' 고지를 의무화하고, 이에 대한 이용자의 이용거부, 설명요구 및 이의제기 권리를 제도화(GDPR, '18.~)

- 신뢰가능한 인공지능의 3대요소(적법성·윤리성·견고성)를 제시('19.)하고, 민간이 신뢰성 등을 자율점검할 수 있는 체크리스트 제작·보급('20.)

○ **[미국]** 신뢰성 확보 **기술개발**에 **국가역량**을 **결집**하는 한편, **주요 기업**을 중심으로 **윤리적 인공지능 실현**을 위한 **자율규제*** **전개**

* 인공지능 개발원칙 마련(구글, MS, 딥마인드 등), 공정성 점검도구 개발·공유(IBM 등)

- **과잉규제 지양**과 **위험 기반 사후규제 기조** 하에 인공지능 **신뢰확보 10대 원칙**(투명성, 공정성 등)을 담은 **규제 가이드라인** 발표('20)

○ **[프랑스]** 기업·시민 등 3천명이 참여한 **숙의적인 공개토론**을 통해 '**인간을 위한 인공지능**' 구현에 필요한 **권고사항** 도출('18)

※ <영국> 5대 윤리규범('18), 설명가능한 인공지능 가이드라인('20), <일본> 인간 중심의 사회원칙('18) 등 추진

□ **[국내] 기술·제도·윤리적 지원**을 통해 **신뢰성 향상 가속화** 필요

○ **[기술]** '설명가능', '공정', '견고' 측면의 **원천기술을 개발**(총 295억원) 중이며, 산업 전반 **신뢰 구현**을 **뒷받침**할 **검증 지원체계** 필요

○ **[제도]** 인공지능 관련 **법·제도의 정비 과제**는 **마련***한 상황으로, **민간 책임성 강화**와 **규제 불확실성**을 **최소화**하는 방향으로 **신속 정비** 필요

* 「인공지능 법·제도·규제 정비 로드맵」('20.12.)을 통해 관계부처 소관 30개 과제 발굴

○ **[윤리]** **「인공지능 윤리기준」**('20.12.)과 함께, IT기업(카카오 등)을 중심으로 추진되는 **민간 중심**의 **윤리 정립**과 **의식 확산**을 **촉진**할 필요

☞ **인공지능**의 **신뢰 확보**를 위해 ①**민간**의 **인공지능 서비스 구현**을 **체계적**으로 **지원**하고, ②**이용자**가 **믿고 안전**하게 **사용**할 수 있도록 **제도**를 **보완**하며, ③**사회 전반**에 건전한 **인공지능 윤리 확산** 추진

Ⅲ 비전 및 목표

비전	"누구나 신뢰할 수 있는 인공지능, 모두가 누릴 수 있는 인공지능 구현" - Trustworthy AI for Everyone -		
목표 (~'25)	책임있는 인공지능 활용 세계 5위	신뢰 있는 사회 세계 10위	안전한 사이버국가 세계 3위
추진 전략	**신뢰 가능한 인공지능 구현 환경 조성**	**안전한 인공지능 활용을 위한 기반 마련**	**사회 전반 건전한 인공지능 의식 확산**
	① 인공지능 제품·서비스 신뢰 확보 체계 마련 ② 민간 신뢰성 확보 지원 ③ 인공지능 신뢰성 원천기술 개발	① 학습용 데이터 신뢰성 제고 ② 고위험 인공지능 신뢰 확보 ③ 인공지능 영향평가 추진 ④ 신뢰 강화 제도 개선	① 인공지능 윤리 교육 강화 ② 주제별 체크리스트 마련 ③ 인공지능 윤리정책 플랫폼 운영

Ⅳ 추진전략 및 과제

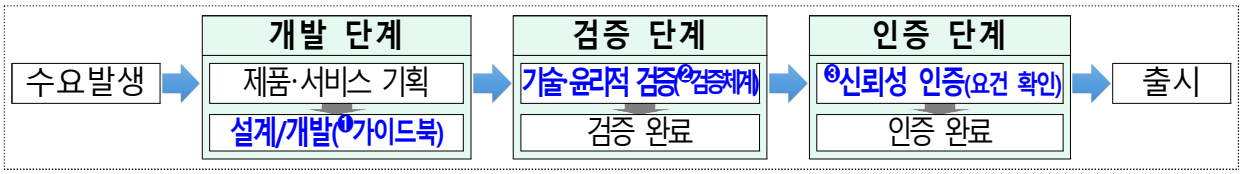
전략1 신뢰 가능한 인공지능 구현 환경 조성

① **[신뢰확보 체계]** 제품·서비스 구현단계(개발→검증→인증)에 따라, '개발 가이드북*', '검증체계**', '인증' 등 신뢰 확보 기준·방법론 제시·지원

* 신뢰 구현에 참조할 수 있도록 관련 **법·제도·윤리·기술적 요구사항** 등을 구체화하여 **종합 제시**

** **가이드북**을 **준수**한 **신뢰성 확보 여부·수준** 등을 **확인·평가** 할 수 있는 **검증절차·항목·방법**

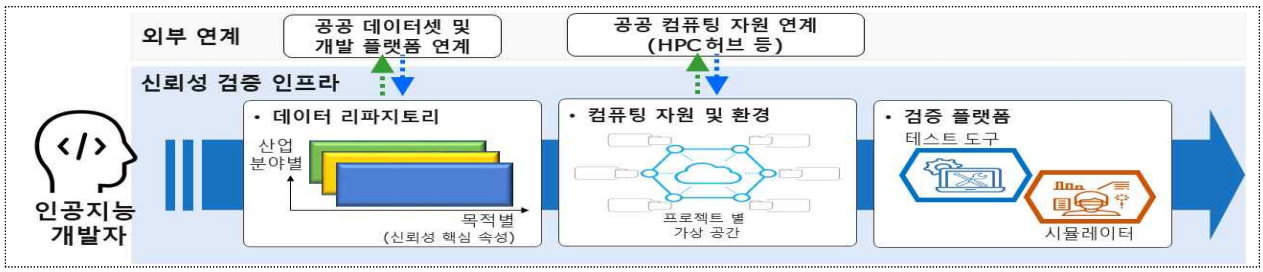
< 인공지능 제품·서비스 구현단계별 지원 흐름도>



② **[민간 신뢰성 확보 지원]** 중소기업 등의 '데이터 확보→알고리즘 학습(구현)→검증'을 **통합 지원**하는 **온라인 플랫폼*** 구축·운영

* 학습용 데이터와 고성능 컴퓨팅 자원을 지원 중인 **'인공지능 허브' 플랫폼**에서 검증체계에 따른 **신뢰 속성별 수준 분석**, **실(實)환경 테스트** 등의 기능을 **추가 개발·연계**하여 **일괄 지원**

< '데이터 확보 → 알고리즘 학습(컴퓨팅) → 검증' One-Stop 지원 플랫폼 구성안 >



③ **[원천기술 개발]** 기존 시스템에 설명가능성 기능을 추가할 수 있고, 인공지능이 스스로 편향성을 진단·제거할 수 있는 기술 등 개발 추진

구분	'설명가능' 분야 기술개발(안)	'공정' 분야 기술개발(안)	'견고' 분야 기술개발(안)
기간/예산	'22~'26년(총 5년) / 총 450억원*	'22~'26년(총 5년) / 총 200억원*	'22~'26년(총 5년) / 기획 추진 중
주요 내용	설명가능성에 대한 고려 없이 **개발된 모델에 전문가 보고서 생성** 수준의 **설명가능성 제공**	**규정 등**으로부터 스스로 **편향 요인을 식별**, **편향 가능성 진단**하고 **최적**의 **제거 방안**을 **판단·적용**	**복합정보**(음성, 영상, 위치 등)를 이용한 인공지능에 대한 공격의 **선제적 대응** 기술

* 예타 통과 금액 기준

전략2 안전한 인공지능 활용을 위한 기반 마련

① **[학습용데이터 신뢰성 제고]** 全 제작공정에서 **민·관**이 공통 **준수**할 **표준 기준**을 **마련***·**확산**하고, **데이터댐 사업**의 **품질 향상**** 추진

* 학습용 데이터 활용 목적에 따라 **신뢰 확보 요구사항**을 **세분화·구체화**하고 **검증지표, 측정방법** 등 제시

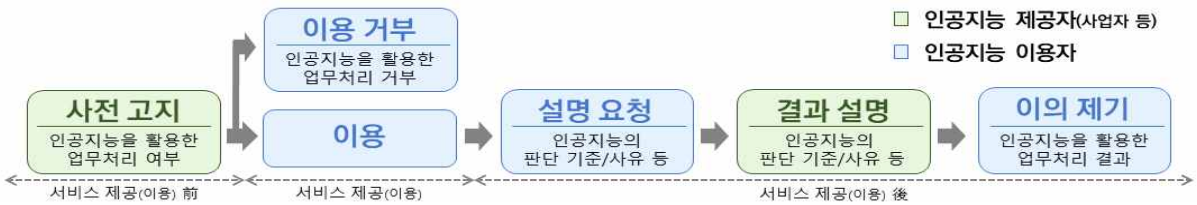
** 법·제도(저작권, 개인정보보호 등) 준수 여부 등 공정별 신뢰성 제고 고려사항을 적용

② **[고위험 인공지능 신뢰 확보]** **잠재적 위험**을 미칠 **인공지능**의 **범주**를 **설정***하고, 서비스 제공前 **해당 인공지능**의 **활용 여부 고지****를 **의무화** 추진

* <사례> EU는 시민의 안전, 기본권에 부정적 영향을 미치는 시스템으로 정의(AI regulation, '21.4)

** 고지 이후 해당 인공지능 기반 서비스에 대한 '이용 거부', 인공지능의 판단 근거에 대한 '결과 설명' 및 이에 대한 '이의제기' 등에 대해서는 다각도로 중장기 검토

< 고위험 인공지능 서비스의 운용/이용단계에 따른 주체별 의무/권리(안) >



※ '자동화된 의사결정'에 대한 동 의무/권리는 △「유럽 개인정보보호법」(GDPR, '18.5~)은 기 제도화, △우리나라는 「개인정보보호법」('21.1.6일 입법예고)에서 제도화 추진 중

③ **[인공지능 영향평가 추진]** **인공지능**이 **국민생활 전반**에 **미치는 영향**을 체계·종합적으로 **분석**하고 **대응**하기 위해 **영향평가 실시**

※「지능정보화기본법」 제56조(지능정보서비스 등의 사회적 영향평가)의 실질적 추진

- 신뢰성 요소(안전성, 투명성 등)를 토대로 분석하여 **인공지능**의 **영향력**을 **종합 분석**하고 **기술·관리적 조치방안**에 대해 **제시·권고 추진**

④ **[신뢰 강화 제도 개선]** 사회 전반에서 활용되는 **인공지능**에 대한 **신뢰 확보, 이용자**의 **생명·신체 보호** 등을 위한 **법·제도 정비 4건*** 추진

* 「인공지능 법·제도·규제 정비 로드맵」('20.12.)에서 발굴한 ①업계 자율적 알고리즘 관리·감독환경, ②플랫폼 알고리즘 공정성·투명성, ③알고리즘 공개기준, ④고위험 기술기준 마련 과제

☞ **민간**의 **자율성**을 **저해하지 않는 선**에서, **글로벌 입법 동향**, **인공지능**의 사회·산업적 **파급력** 및 **기술발전 수준** 등을 다각도로 면밀히 고려하여 **제도 개선**의 **범위**와 **대상**을 **명확**히 하고 **불확실성**을 **제거**

전략3 사회 전반 건전한 인공지능 의식 확산

① **[윤리교육 강화]** 인공지능 윤리교육 총론과 교육과정*을 개발하고, 이를 토대로 **시민, 연구·개발자, 학생** 등 대상 **윤리 교육 실시**

* 인공지능이 사회에 미치는 영향, 인간-인공지능간 상호작용 등 사회·인문학적 관점과 윤리 기준의 사회 실천의 필요성을 인식할 수 있는 내용으로 포괄적으로 구성

② **[주체별 체크리스트 마련]** 인공지능 윤리 실천지침으로 **연구·개발자, 이용자** 등이 **윤리 준수 여부**를 **자율점검**을 할 수 있는 **체크리스트** 보급

- **기술발전 양상**을 **반영**하고, **他분야 자율점검표**와도 **체계성·정합성**을 **유지**(「인공지능 윤리기준」이 기본원칙 지위)하여, **실천가능성 제고**

③ **[윤리 정책플랫폼 운영]** 학계·기업·시민단체·공공 등 **다양한 사회 구성원**이 **참여**하여 **윤리**에 대한 **토의·의견수렴**하는 **공론의장** 운영

※ <사례> EU 집행위는 인공지능 분야 4천여명 이상의 다양한 이해관계자 참여하는 온라인 시민 의견수렴 플랫폼인 "The European AI Alliance(유럽 인공지능 연합)" 운영('18.~)

V 추진 체계

○ 관계부처·민간과 적극적으로 소통하며 체계적 추진

본 문

순 서

Ⅰ. 추진배경

◇ "AI Everywhere" 인공지능 시대 본격화

□ **4차 산업혁명**을 견인하는 **범용기술**인 **인공지능**은 **데이터와 결합**하여 **경제·사회 全분야**에서 **디지털전환**(Digital Transformation)을 촉진

※ 미국 스탠포드대학 앤드류 응(Andrew Ng) 교수는 "인공지능은 새로운 전기다(AI is the new electricity)"라 언급하며, **인공지능**이 가진 **범용기술**로서의 **가치**를 강조

□ 우리 사회는 「디지털 뉴딜」('20.7.)의 추진에 따라, **인공지능 활용**이 **양적·질적**으로 **확대**되고, 각 분야에서의 **다양한 혁신 사례***도 **탄생**

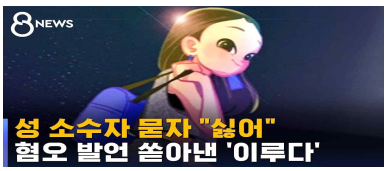
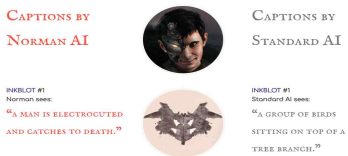
* 자율주행 버스 실증(세종시), 인공지능 용광로(포스코), 인공지능 의사(닥터앤서) 등

○ 특히, 코로나19 대응과정에서 진단키트 개발, 역학조사·모니터링, 진단보조 등에 기여하며 **혁신 기술**로서의 **가치**와 **잠재력**을 확인

◇ 하지만, 인공지능 확산에 따라 예상치 못한 사회적 이슈도 등장

□ 인공지능이 우리 일상생활과 밀접해지면서 **인간 사회**의 다양한 **사회·윤리적 문제**가 **인공지능**에 **투영·남용**되는 사례가 지속 **발생**

< 인공지능이 관련된 주요 사회적 이슈 사례 >

인공지능 챗봇 '이루다'('21.1)	MIT 개발, '사이코패스 AI'('18.6)	오바마 前대통령 딥페이크('18.7)
		
성소수자 혐오발언, 성적인 대화 등으로 출시 1달여 만에 서비스 중단	반사회·반인륜적 데이터로 훈련된 부정·편향적 인공지능 공개	합성된 얼굴과 목소리로 트럼프 前대통령을 모욕하는 영상 유포

□ 인공지능 확산의 이면으로, **인공지능**의 **활용 과정**에서 발생할 수 있는 **위험, 부작용*** 등에 대한 **사회적 관심**과 **우려 대두**

* 개인정보 유출, 사이버 재난, 악성코드, 인공지능 의존도 증가, 저작권 문제 등

※ <참고 : 빌게이츠(美 Microsoft 고문) 발언> "처음에는 인공지능이 우리를 대신해 많은 일을 하겠지만 단 몇 십 년 후에는 심각한 걱정거리가 될 수도 있다"('15.)

◇ 신뢰 가능한 인공지능 구현은 "인공지능 강국" 실현의 대전제

□ 글로벌 주요국은 **인공지능**에 대한 **신뢰 확보**를 인공지능의 **사회적·산업적 수용**과 **발전**의 **전제요소**로 **인식**하고 **인공지능 정책**을 **추진**

○ 국가 인공지능 전략에 '안전한 인공지능(Safe AI)', '신뢰할 수 있는 인공지능(Trustworthy AI)' 등을 명시하고 규범, 기술기준 등을 마련 중

< 주요국 인공지능 신뢰 확보를 위한 주요 정책 추진 현황 >

EU	■ 인공지능법 제안('21), 인공지능백서('20) ■ 신뢰할 수 있는 인공지능 윤리 가이드라인('19)	**미국**	■ 인공지능 규제 가이드라인('20) ■ 인공지능 및 알고리즘 이용지침('20)
영국	■ 설명 가능한 인공지능 가이드라인('20) ■ 공공분야 인공지능 활용 지침('19)	**일본**	■ 인간중심의 인공지능 사회 원칙('18) ■ 인공지능 개발 가이드라인('17)

○ 기업, 대학 등 **민간 차원**에서도 인공지능의 **신뢰**를 위한 **원칙**, **신뢰·안전 보장**을 위한 **기술·방법** 등에 대한 **연구***를 본격 착수

* <사례: IBM 社> 신뢰 가능한 인공지능(Trusting AI) 구현을 위한 공정성, 가치정렬, 견고성, 설명가능성, 투명성 및 책임성 등 5대 원칙 정의 및 측정 도구 개발

□ 우리 역시 **사람**이 **중심**이 되는 **'인공지능 강국'**의 **조속한 실현**을 위해 **신뢰할 수 있는 인공지능**에 대한 **정책 가속화 필요**

< 대한민국 인공지능을 만나다, '20.11.25. >

"우리는 **인공지능**이 가져올 편리함과 동시에 사람의 소외를 초래할지도 모를 **어두운 측면**도 무겁게 **고민**해야 합니다. 경제적 가치와 함께 **사람 중심**의 **가치**의 **중요성**을 **생각**하며 미래를 설계해야 할 것입니다."

○ **「인공지능 국가전략」**('19.12.)과 인공지능 시대에 우리 사회가 나아가야할 방향을 제시한 **「인공지능 윤리기준」**('20.12.)을 토대로,

○ 민간*과 함께 **인공지능**의 **혜택**을 **극대화**하면서도, **개발**과 **활용 전반**에서 **위험**을 **최소화**하기 위한 **사회적 신뢰**(Trust) **확보** 추진

* 카카오(인공지능 윤리헌장, '18), 네이버(인공지능 윤리준칙, '21) 등 자정 노력 중

Ⅱ. 인공지능 신뢰성의 개념 및 필요성

◇ (개념) 인공지능 윤리기준 실현과 수용성 향상을 위한 핵심가치

□ '**인공지능 신뢰성**'은 인공지능이 **내포한 위험**과 **기술적 한계***를 **해결**하고, 활용·확산과정에서의 **부작용**을 **방지**하기 위한 **가치 기준**

* 귀납적 작동원리에 따른 불투명성, 학습에 활용한 데이터에 따른 편향성 등

○ 각 국에서도 **인공지능 윤리**의 **실현**을 위한 **핵심요소**로서 **신뢰성**을 **강조**하고, **안전**, **설명가능**, **투명**, **견고**, **공정** 등을 **속성**으로 포함

< 인공지능 신뢰성(Trustworthiness)의 주요 핵심요소 및 의미 >

속성	주요 의미
안전 (Safety)	■ **인공지능의 판단·예측 결과**로 인한 시스템 동작과 기능 수행이 **사람**과 **환경**에 **악영향**을 **미치지 않도록 예방**할 수 있는 상태
설명가능 (Explainability)	■ 인공지능의 **판단·예측**의 **근거**와 **결과**에 이르는 **과정**이 **사람이 이해 가능한 방식**으로 **제시**되거나, 문제 발생 시 **결과 도출과정**의 **분석**이 **가능**한 상태
투명 (Transparency)	■ **인공지능**의 판단·예측 등 **작동과정**과 이를 구현하기 위한 **구성 요소**에 있어 **이용자**가 **인지**하고 **확인·검사**가 **가능**한 상태
견고 (Robustness)	■ 인공지능이 **외부**의 **간섭** 및 **극한적인 운영 환경**에서도 사용자가 **의도한 수준의 성능** 및 **기능**을 **유지**하는 상태
공정 (Fairness)	■ 인공지능이 **데이터**를 **처리**하는 **과정**에서 **특정 그룹**에 대한 **차별**이나 **편향**을 **포함하는 결론**을 **도출하지 않도록** 하는 기능성

※ 프라이버시(Privacy), 지속가능성(Sustainability) 등도 핵심요소 중 하나로서 다양하게 논의 중인 상황

[참고 : 주요 기관에서 논의 중인 인공지능 신뢰성 개념]

■ **(국제표준화기구, ISO)** 신뢰성의 세부 속성으로 가용성, 회복탄력성, 보안성, 안전성, 프라이버시, 책임성, 투명성, 통합성 등 제시(ISO/IEC TR 24028, '20.)

■ **(경제협력개발기구, OECD)** 지속 가능한 사회와 인간 중심의 가치에 부합하고 투명성 및 설명가능성, 견고성 및 안전성을 갖춘 인공지능(Recommendation of the Council on AI,'19.)

■ **(美 국립표준연구소, NIST)** 인공지능 신뢰성과 관련된 특성으로 정확성, 탄력성, 객관성, 보안성, 설명가능성, 안전성 등을 제시(U.S. leadership in AI, '19)

■ **(유럽위원회, EC)** 인공지능은 활용 및 동작이 합법적이어야 하며, 윤리적이어야 하고 기술적·사회적으로 견고해야 함(Ethics guidelines for trustworthy AI, '19.)

◇ [필요성] 사회·윤리적 문제를 야기할 수 있는 인공지능에 대한 보완

□ **(기술적 특성)** 인공지능 구현에 주로 사용되는 '**기계학습**(Machine Learning)' 방식은 시스템이 **데이터**를 **기반**으로 **스스로 반복 학습**하며 **고도화**하는 것이 특징

<인공지능 구현 과정>

○ 이에, **인공지능**은 **학습 데이터**에 따른 **편향**(偏向), 불투명한 결과 도출과정에 따른 **의도되지 않은 차별 발생** 등의 잠재위험 존재

< 인공지능의 기술적 특성으로 발생된 주요 사회·윤리적 이슈 사례 >

▪ **(데이터 편향성)** 아마존의 인공지능 기반 채용시스템이 개발자, 기술 직군에 **대부분 남성만**을 **추천**하는 문제가 발생함에 따라 아마존에서 동 시스템의 사용을 폐기('18.10.)

▪ **(알고리즘 차별)** 미 20여개 주 법원에서 사용하던 인공지능 기반 범죄 예측 프로그램인 'COMPAS'의 **재범률 예측**에서 **흑인 범죄자**의 **재범 가능성**을 **백인보다 2배 이상 높게 예측**하는 편향 발견('18.1.)

□ **(활용적 특성)** 우리 삶 전반(의료 등)에서 **활용**되고 **파급력** 큰 특징

○ **부주의, 오·남용/악용**에 따라 **사생활 침해, 사회·경제적 피해**뿐 아니라 예상할 수 없는 **대규모 복합문제**를 **유발**할 잠재위험 존재

< 인공지능의 활용적 특성으로 발생된 주요 사회·윤리적 이슈 사례 >

▪ **(기술 오·남용)** 유럽 한 에너지기업의 CEO는 영국 범죄자들이 **인공지능**을 **활용**해 정교하게 **만든** 모회사 CEO의 **가짜음성**에 **속아** 22만 유로를 송금하는 피해('19.9.)

▪ **(사생활 침해)** 아마존 '알렉사', 구글 '구글 어시스턴트', 애플 '시리' 등의 **인공지능 스피커**로 **수집**된 **음성정보**를 제3의 외부업체가 **청취**하는 것으로 밝혀져 논란('19.9.)

▪ **(복합 문제)** '스캐터랩'이 출시('20.12.)한 인공지능 챗봇 '이루다'는 △**데이터 구축·학습과정**에서의 **개인정보 침해 가능성**, △**데이터·알고리즘**의 **편향성**, △이용자의 **챗봇 성희롱 대화** 등으로 큰 사회적 이슈가 되어 결국 1달여 만에 자체 서비스 중단

◈ 인공지능의 **기술적 한계 극복**과 함께, 오·남용 등에 따른 **잠재 위험**의 **예방**을 위해 **제도 보완**과 **윤리 의식**의 **확산**이 필요

참고1 「인공지능 윤리기준」('20.12) 과 인공지능 신뢰성

□ 「인공지능 윤리기준」은 '인간성을 위한 인공지능'(AI for Humanity)을 위해 모든 사회구성원이 지켜야 할 **3대 원칙**과 **10대 요건**을 제시

< 「인공지능 윤리기준」 주요내용 >

- **[최고 가치]** 윤리기준이 지향하는 최고 가치를 **'인간성(Humanity)'**으로 **설정**
- **[3대 기본원칙]** 인공지능의 개발 및 활용 과정에서의 '인간성(Humanity)' 구현을 위한 **❶인간의 존엄성 원칙**, **❷사회의 공공선 원칙**, **❸기술의 합목적성 원칙**
- **[10대 핵심요건] 3대 기본원칙의 실천·이행**을 위한 **①인권 보장, ②프라이버시 보호, ③다양성 존중, ④침해금지, ⑤공공성, ⑥연대성, ⑦데이터관리, ⑧책임성, ⑨안전성, ⑩투명성** 요건

□ **인공지능 신뢰성**(Trustworthiness)은 **인공지능 윤리**의 **실천**과 **이용자**의 **인공지능 수용성 향상**을 위한 **핵심 요구사항**으로, 안전, 설명가능, 투명, 견고, 공정 등을 포함하는 광의의 개념

○ 인공지능 **개발·활용·확산**에서 **신뢰성**의 **구현 기준**과 **방안**을 **정립·추진**함으로써 **인공지능**에 대한 **국민**의 **수용성**을 **향상**

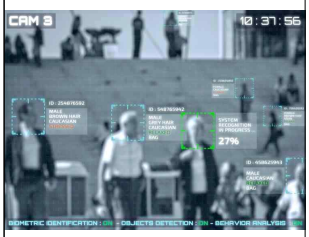
○ 신뢰성 향상을 위한 기술적, 제도적, 법적 정비 및 지원 정책이 요구됨

참고2 최근 발생한 인공지능 신뢰성 관련 주요 이슈 사례

[데이터 편향] 英 정부, 차별 야기한 알고리즘 기반 성적 산출 시스템 철회 결정

- 영국은 코로나로 대학입학시험을 취소하고, △담당교사 평가, △출신 학교의 과거 성적분포를 토대로 인공지능으로 학생의 예상성적 산출
- 공립학교와 빈곤지역 학생 성적이 사립학교의 부유층 학생 대비 저조하게 산정되어, 정부는 인공지능 시스템의 철회 결정

[인권 침해] 안면인식 기술, 법집행 기관의 감시 강화 및 사생활 침해 도구로 악용 우려



- 美 경찰 당국은 범죄자 식별, 시민 감시 등을 위해 안면인식 인공지능 기술을 활용
- 범죄자 오인, 인종 차별 등 기술적 한계와 오남용 문제로 아마존, IBM, MS 등은 법집행 기관에 관련 기술 판매 중단 결정('20.6.)

[정치적 혼란 야기] 페이스북 회원정보, 美 대선의 정치적 선동에 활용

- 케임브리지 애널리티카는 페이스북 회원정보를 프로파일링 하여 2016년 대선에서 트럼프에 유리하도록 선거 개입
- 美 연방거래위원회(FTC)는 페이스북의 고객 데이터 부실 관리에 대해 50억 달러의 벌금을 부과('19.7.)

[사회적 편견 악화] 마이너리티 리포트 현실화 논문에 인공지능 연구자 집단 반발

- 美 Harrisburg 과기대 연구진은 안면인식 기반의 범죄예측 연구결과를 獨 Springer 발간 학술지에 게재 예정
- 약 2,400명의 인공지능 연구자는 신경망 학습을 통한 범죄 예측은 사회적 편견을 악화시킨다며 철회 요구 서한 공개('20.6.)

[안전성 미흡] 테슬라 자율주행 중 전복 트럭을 인지 못해 사고 발생

- 대만 고속도로에서 자율주행모드로 주행하던 테슬라 차량이 전복된 트럭을 인지 못해 정면충돌('20.6.)
- 테슬라의 자율주행 기능은 특정한 색으로 도색한 차량 등을 인지하지 못해 지속적으로 사고가 발생

Ⅲ. 정책 환경 분석

1 글로벌 동향

◇ (EU) 인공지능 신뢰성 제고를 위한 '제도', '윤리' 정립을 주도

□ **(제도)** 「인공지능 법안」('21.4.)을 통해 **세계 최초**로 인공지능 신뢰 확보를 위한 **위험기반**(Risk-based)의 **인공지능 시스템 규제체계*** 제시

* △금지(Unacceptable risk), △고위험(High-risk), △제한된 위험(Limited-risk), △최소 위험(Minimal-risk)

○ 안전 등에 부정적 영향을 끼칠 **'고위험 인공지능'**을 **중심**으로 **규제**하고 **공급자 등**에게 **사용자 정보제공** 등의 **의무***를 **부과**

* △적절한 위험관리시스템 운용, △고품질의 데이터셋 관리, △당국에 상세한 문서 제공, △결과추적을 위한 로그 활동의 기록 및 관리, △사람의 감독 등

- 제품 출시 전 의무 이행 여부 확인을 위한 **적합성 평가** 및 **인증** 실시

○ 한편, GDPR(개인정보보호법)을 통해 **자동화된 의사결정**에 대한 △**사전 고지**, △**이용 거부**, △결과 설명요구, △**이의제기 권리 보장**을 제도화

< GDPR 상 자동화된 의사결정에 대한 권리 흐름도 >

□ **(윤리)** **신뢰 가능**한 **인공지능**의 **3대 구성요소**(적법성·윤리성·견고성)을 포함한 가이드라인('19)을 시작으로, **지침, 교육, 행동강령** 등 개발

○ 민간이 인공지능 개발 全과정에서 **신뢰성 등 윤리 이슈**를 **자체 점검**할 수 있는 **평가목록***(**체크리스트**) 배포('20.7)

* 자율성 및 감독, 투명성, 개인정보보호 등 7대 분야, 총 146개 항목 구성

□ **(기술)** 신뢰 관련 기술에 대한 연구개발*('20~'24, 총 443억 원 규모)도 병행하며, 신뢰할 수 있는 인공지능 기술 개발 주도권 확보 노력

* 안전하고 재사용 가능한 인공지능 플랫폼('21~'23, €8M), 사생활 보호를 위한 연합학습('18~'21, €4.4M) 등

◇ (미국) 신뢰 가능한 인공지능 '기술' 개발에 집중

□ **(기술) 국가 인공지능 R&D 전략**으로서, 기술적으로 **안전한 인공지능 개발** 등을 채택*하고 관련 기술 확보에 **민관**의 **역량**을 **결집**

* 국가 인공지능 R&D 전략 보고서(美국가과학기술위 인공지능특별위, '19.)

○ 특히, **글로벌 빅테크 기업**(구글, 아마존 등)은 자체 개발과 함께, 우수 기술 **연구자금 지원***을 병행하며 **신뢰성 기술 확보**에 노력

* 책임 있는 인공지능 개발을 위해 구글, 아마존, 마이크로소프트 등이 조직한 'Partnership on AI'는 전미과학재단(NSF)에 연구비 450만 달러 지원('18.)

□ **(제도·윤리) 과잉규제 지양과 위험기반 사후규제** 기조 하에 인공지능 **신뢰확보 10대 원칙***을 담은 연방정부 **규제 가이드라인** 발표('20.1.)

* ①대중의 신뢰, ②시민 참여, ③과학적 무결성과 정보 품질, ④위험 측정 및 관리, ⑤비용편익, ⑥유연성, ⑦공정성과 차별금지, ⑧정보공개와 투명성, ⑨안전과 보안, ⑩기관 간 협력

○ 주요 **빅테크 기업**을 중심으로 윤리적 **인공지능 개발원칙 마련***과 **공정성 점검 도구**** 개발·공유 등 **자율적 정화 분위기** 조성 중

* DeepMind 윤리와 사회원칙('17.10), Google 인공지능 원칙('18.6), Microsoft 인공지능 원칙('18.11) 등

** IBM(AI Fairness 360), Google(What-if Tool), Microsoft(Fair Learn) 등

◇ (영국, 프랑스, 일본) 신뢰성 확보를 위한 '윤리' 정립 노력 추진

□ **(영국) 5대 윤리규범***('18.4.), 공공부문 안전한 인공지능 활용을 위한 지침('19.6.), 설명 가능한 인공지능 가이드라인('20.5.) 등 수립

* ①인간에 이롭게 활용, ②공정성, ③개인정보보호, ④인공지능의 파급효과에 대해 국민이 교육 받을 권리, ⑤인공지능의 안전성 확보

□ **(프랑스)** 기업·시민 등 약 **3천명**의 **숙의적인 공개토론**을 통해 **'인간을 위한 인공지능'** 구현에 필요한 **권고사항 도출**(「인공지능윤리보고서」, '18.8)

□ **(일본)** 인공지능과 관련된 모든 이해관계자들이 유의해야 할 7대 기본 원칙을 담은 **「인간 중심의 인공지능 사회 원칙」**('18.3.) 발표

참고3 주요국 인공지능 신뢰성 주요 정책 내용

구분	주요 내용
 EU	■ **신뢰할 수 있는 인공지능 윤리 가이드라인**('19.4.) **인공지능**이 **준수**해야 하는 **기본원칙·가치***를 **제시**하고, 이를 바탕으로 **신뢰할 수 있는 인공지능 구현방안****을 도출, **평가목록(체크리스트) 방향 제시** * (인공지능 윤리 4원칙) ①자율성 ②무해성 ③공정성 ④설명가능성 ** (신뢰할 수 있는 인공지능 실현을 위한 주요 요건) ①인간주체성·자율성, ②견고성과 안전성, ③프라이버시와 데이터 거버넌스, ④투명성, ⑤다양성·차별금지·공정성, ⑥사회적·환경적 웰빙, ⑦책임성 ■ **신뢰할 수 있는 인공지능 평가목록**('20.7) 인공지능 개발 과정에서 **신뢰할 수 있는 인공지능 구현**하고 있는지 **스스로 판단**할 수 있도록 하는 **평가목록(체크리스트) 제공**(7개 분야* 146개 항목**) * (평가 분야) ①인간의 대리·감독, ②기술적 견고성·안전성, ③프라이버시·데이터 거버넌스, ④투명성, ⑤다양성·차별금지·공정성, ⑥사회적·환경적 웰빙, ⑦책임성 ** (예) 필요할 때 작업을 안전하게 중단하기 위한 '중지 버튼' 또는 절차를 보장했는가?, 최종 사용자에게 보안 적용 및 업데이트 기간을 알려주었는가? 등
 미국	■ **인공지능 규제 가이드라인**('20.1) 인공지능 규제 수립 시 **과잉규제를 지양**하고 **규제의 영향을 면밀히 검토**한다는 전제하에 **인공지능 신뢰성을 확보하기 위한 10대 원칙***을 제시 * ①대중의 신뢰, ②시민 참여, ③과학적 무결성과 정보 품질, ④위험 측정 및 관리, ⑤비용편익, ⑥유연성, ⑦공정성과 차별금지, ⑧정보공개와 투명성, ⑨안전과 보안, ⑩기관 간 협력
 영국	■ **공공분야 인공지능 활용 지침**('19.6.) **공공분야의 인공지능 활용·확산**을 **목적**으로 공개한 실무 수준의 지침으로 인공지능의 **윤리적**이고 **안전한 활용***을 위한 **지침** 포함 * (고려요인) 데이터품질, 공정성, 개인정보보호, 투명성, 설명 가능성, 비용 등 제시
 프랑스	■ **인공지능 윤리문제 보고서**('18.8) **인공지능 윤리** 문제에 대해 **시민사회**의 **집단적 인식**을 높이기 위해 윤리에 대한 **주요 주제***와 윤리 정책 수립을 위한 **권고사항**** 제시 * (주요주제) 데이터 편향성, 알고리즘에 의한 차별 및 배제 등 ** (권고사항) 개발자·전문가·시민 등 모든 이해관계자 교육, 이해할 수 있는 알고리즘 시스템 만들기, 인공지능 비즈니스 분야의 윤리 강화
 일본	■ **인간중심의 인공지능 사회 원칙**('18) **적극적인 인공지능 사회**를 **구현**하기 위해서 **모든 이해 관계자**가 유의해야 할 **7대 기본원칙***을 수립 * (7대 기본원칙) ①인간중심, ②교육·교양, ③개인정보보호, ④보안 확대, ⑤공정 경쟁, ⑥공정성·책임성·투명성, ⑦혁신

2 국내 현황 및 대응방향

◇ [기술] 인공지능 신뢰성 확보를 위한 기술 역량 확보 시급

□ 미국, 유럽 등은 **인공지능 신뢰성** 관련 **기술개발**에 투자 확대 중*

* <美 DARPA> 투명성·견고성·안전성 확보를 위한 차세대 인공지능 R&D 과제('17~'21, 1,830억원), <美 NIST> 인공지능 신뢰성 확보를 위한 연구개발('21, 280억원), <EU> 'HORIZON 2020 프로젝트' 內 투명성·편향성 등 신뢰성 R&D('18~'24, 443억원)

○ 우리는 현재 인공지능의 **설명가능, 공정**(편향성), **견고** 측면에서 신뢰 확보를 위한 **기술개발**을 추진 중(총 295억원)

< 국내 인공지능 신뢰성 확보를 위한 기술개발 과제 주요내용>

속성	설명
설명가능	■ **인공지능**이 **의사결정**한 **이유**(판단 기준/사유)를 **설명**할 수 있도록 하는 학습·추론 설계 기술 개발('17~'21, 총 184.7억원)
공정 (편향성)	■ **인공지능**(알고리즘 모델) 및 **학습 데이터**의 **편향성** '분석 → 탐지 → 완화 → 제거' 기술 개발('19~'22, 총 50.1억원)
견고	■ **인공지능 기만공격** 사례 분석 및 사전 감지 기술('18~'20, 총 13.65억원), 인공지능 보안 **역기능·취약점 자동탐지** 기술 개발('20~'27, 총 46.5억원) 등

□ 신뢰 가능한 인공지능 **개발** 등에 대한 **논의**가 **아직 활성화**되지 **못해**, 인공지능이 수행하는 의사결정 등에 대해 우려가 있는 상황

○ **글로벌 표준화 기구**(ISO, IEC 등)에서 '17년부터 신뢰, 윤리에 대한 **논의**를 **진행 중**이며, **우리**도 **표준화 논의 착수**('19~)

○ 기업을 중심으로 **인공지능**이 **잘못된 의사결정** 또는 **예상치 못한 결과**를 **초래***하여 손해·고객이탈 등이 발생 할 수 있다는 우려 有

* 인공지능 관련 기업 우려사항 조사(KDI, '20) : 인공지능의 △의사결정·행동의 법적책임(23.1%), △잘못된 의사결정(21.6%), △보안취약성(19.0%), △실수로 고객 신뢰훼손(12.0%) 순

◈ [상황진단] 신뢰 가능한 인공지능 구현을 위한 기술 기반 취약

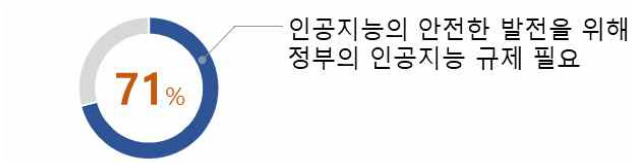
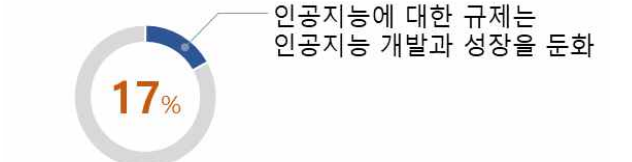
☞ **(대응방향) 인공지능 신뢰성**(공정, 견고 등) **원천기술**의 **개발**과 제품·서비스 **개발·활용** 全과정에서의 **신뢰성 확보 체계적 지원** 추진

◇ [제도] 책임 있는 인공지능 활용을 위한 제도 기반 마련 필요

□ 인공지능 기술의 발전 속도와 방향을 고려한 **선제적**이고 **합리적**인 **법·제도적 체계**에 대한 **고민** 및 **개선**이 **필요**

○ 기업 등은 **인공지능** 기반 **新서비스**의 **규제 적용여부**, 인공지능의 **의사결정**에 대한 **책임소재** 등에 대한 명확한 법제도 정비 희망

< 인공지능 관련 법·제도에 대한 인식 조사(서울시 미래경영청년네트워크, '20.12.) **>**

인공지능 규제의 필요성	인공지능 규제와 기술발전의 관계
	
☞ 안전한 발전을 위해 정부 규제 필요하다고 인식	☞ 규제가 기술 발전을 저해하지 않는다고 인식

□ 지난해 범정부 차원에서 **「인공지능 법·제도·규제 정비 로드맵*」**('20.12.)을 **마련**하여 법·제도 **정비방향**과 **우선 과제**는 **도출**한 상황

* 과기정통부, 산업부, 개보위, 공정위 등 관계부처 소관 **30개 과제**로 **구성**

< 「인공지능 법·제도·규제 정비 로드맵」 주요내용 >

산업 진흥과 활용 기반 강화 (18과제)	역기능 방지 (12과제)
△데이터기본법 제정, △저작권법 개정, △인공지능 법인격 부여, △자율적 알고리즘 관리 가이드라인 마련, △인공지능 행정 근거 마련 등	△디지털포용법 제정, △알고리즘 공정성·투명성 확보, △인공지능 윤리기준 마련, △이상거래 대응체계 강화 등
* 법령(19건), 지침·가이드라인(2건) * 단기(11건, '21~'22), 장기(10건, '23~)	* 법령(9건), 지침·가이드라인(4건), 제도운영개선(4건) * 단기(15건, '21~'22), 장기(2건, '23~)

◈ [상황진단] 제도적 불확실성이 인공지능 확산의 리스크로 작용

☞ **(대응방향) 개별 법제**의 신속한 **정비**와 함께, **민간**의 **책임성**을 **강화**하고, **불확실성**을 **최소화**하는 방향으로 **제도적 보완** 추진

◇ [윤리] 사회 전반적 인공지능 윤리 의식 저조

□ 국가 차원의 「인공지능 윤리기준」('20.12.) 마련과 함께, **IT기업**을 중심으로 **민간 차원의 윤리원칙 수립***, **교육 등 자정 노력**이 병행 중

* 카카오 알고리즘 헌장('18.1), 네이버 인공지능 윤리('21.2), 삼성전자 인공지능 윤리 제정 착수('20.12.)

< 「인공지능 윤리기준」 상 인공지능 윤리 3대 원칙 >

- **[인간 존엄성 원칙]** 인공지능의 개발 등은 인간의 **생명과 건강에 해가 되지 않도록** 안전성·견고성 기반으로 이루어질 것
- **[사회의 공공선 원칙]** **취약계층**의 **지능정보사회 접근성**을 **보장**하고 **인류**의 **보편적 복지**를 **향상**시키는 **방향**으로 **인공지능을 개발·활용**할 것
- **[기술의 합목적성 원칙]** **인공지능**은 인간에게 도움이 되어야 하며 **인류의 삶**과 **번영**을 위하여 **공헌**할 것

□ **사회 전반**으로의 **인공지능 윤리**를 **확산**하고 스타트업, 일반시민 등의 **윤리인식**을 **제고**하기 위한 **정책적 노력 가속화 필요**

○ 정부의 **인공지능 준비도**('20, Oxford Insights)는 종합적으로는 높으나(194개국 中 7위), **책임있는 인공지능 활용*** **측면**은 일부 **개선 필요**

* 34개국 中 21위 평가 / '20년 신규 추가된 지표(Responsible Use)로 국가별 포용성, 투명성 등을 평가하며, 우리는 일본, 싱가포르 등에 뒤쳐진 21위 수준으로 평가

○ 인력·재원이 부족한 **중소기업·스타트업**의 자체적 **윤리원칙 수립**과 **준수***에는 **한계**가 있고, 국민의 인공지능에 대한 신뢰도**도 아직 저조

* 챗봇 '이루다'를 개발한 스타트업은 소프트뱅크로부터 50억원을 유치하였으며, 대표는 포브스(Forbes)紙 2030파워리더에 선정되었던 인공지능 전문기업

** 일반시민 456명 중 55명(12%)만 개인적으로 중요한 결정을 인공지능에게 맡길 수 있다고 응답(서울시 미래경영청년네트워크, '20.12.)

◈ [상황진단] 인공지능 윤리에 대한 인지도와 실천 방안 미흡

☞ **(대응방향)** 인공지능 **윤리교육 강화**, 윤리 체크리스트 등 **주체별 실천방안** 마련, **국민 소통 확대** 등 **사회 전반 윤리 확산** 추진

Ⅳ. 비전 및 목표

비전

"누구나 신뢰할 수 있는 인공지능,
모두가 누릴 수 있는 인공지능 구현"

- Trustworthy AI for Everyone -

목표 (~'25)

책임있는 인공지능 활용	신뢰 있는 사회	안전한 사이버국가
세계 5위	세계 10위	세계 3위
※책임있는 인공지능 활용(Oxford Insights) 현재 34개국 중 21위('20)	※정부에 대한 신뢰수준(OECD) 현재 43개국 중 28위('19)	※사이버보안지수(ITU) 현재 175개국 중 15위('18)

추진 전략 및 과제

신뢰 가능한 인공지능 구현 환경 조성

① 인공지능 제품·서비스 신뢰 확보 체계 마련
② 민간 신뢰성 확보 지원
③ 인공지능 신뢰성 원천기술 개발

안전한 인공지능 활용을 위한 기반 마련

① 학습용 데이터 신뢰성 제고
② 고위험 인공지능에 대한 신뢰 확보
③ 인공지능 영향평가 추진
④ 사회 전반 신뢰 강화 제도 개선

사회 전반 건전한 인공지능 의식 확산

① 인공지능 윤리 교육 강화
② 주체별 체크리스트 마련
③ 인공지능 윤리 정책 플랫폼 운영

Ⅴ. 추진전략 및 과제

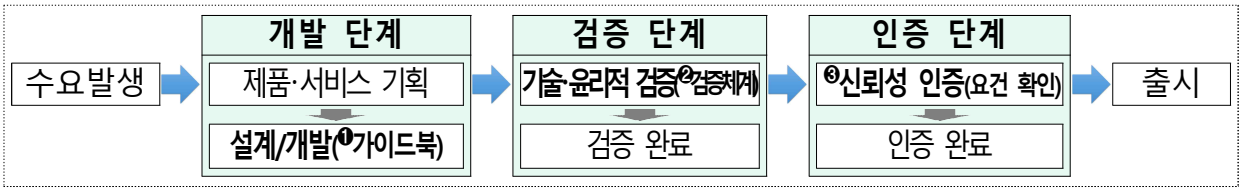
1 신뢰 가능한 인공지능 구현 환경 조성

◇ 세계를 선도하는 **인공지능 신뢰성 구현**을 위해 표준 마련 등 개발 전주기 맞춤 지원, 원천기술 개발 추진

1. 인공지능 제품·서비스에 대한 신뢰 확보 체계 마련

□ **인공지능 제품·서비스 구현단계**(개발→ 검증→ 인증)에 따라, **신뢰성 확보**를 위해 고려해야 할 **기준·방법론**을 **제시**하고 민간 지원 추진

< 인공지능 제품·서비스 구현 과정 >



□ **(❶개발 가이드북)** 인공지능 **개발 단계**에서 **기업, 개발자**들이 **신뢰성 확보**를 위해 **참조**할 수 있는 **'가이드북' 제작·보급**

○ **인공지능 원칙 등***을 기반으로 **핵심속성**(예: 안전성)을 **도출**하고 준수해야 할 **기술적·윤리적 요구사항**을 구체화하여 **종합**** **제시**

* (국내) '인공지능 윤리기준'('20.12.), '개인정보보호 자율점검표(개보위, '21.)' 등
(국외) 'OECD 5원칙'('19.5.), EU 개인정보보호법(GDPR) 및 인공지능 법안 등

** 개발·연구자 대상의 '인공지능 윤리', '개인정보보호', '기술적 속성'을 모두 포함

< 예시 : 안전성 속성 관련 기술적 요구사항 도출 과정 >

OECD의 5원칙	핵심속성	기술적 요구사항(안전성)
■보안과 **안전성**	**안전성**	**❶ 안전한 학습 결과 영역 정의** - 시스템의 안전한 동작 및 기능 정의 - 데이터 학습 범위 설정·제약 등 **❷ 안전 기능 및 동작 보장** - 잘못된 판단을 한 경우 보완책 및 방지 기능 정의 및 구현 ⋮
■포용성장, 지속가능한 발전	설명가능성	
■인간중심 가치와 공정성	견고성	
■투명성 및 설명가능성	보안성	
■책임성	공정성	
	⋮	

○ **국내·외 신뢰성 표준 논의**와 **연계**하여 **고도화·표준화**하고, 향후 의료, 제조 등 **활용 분야별** 특화된 **신뢰 구현** 및 **적용 가이드** 마련

□ (❷검증체계) **'가이드북'**을 **준수**한 **신뢰성 확보 여부, 수준** 등을 체계적으로 **확인·평가**하기 위한 **'검증체계' 개발·보급**

○ 기업, 제3자 등이 **개발된 인공지능**의 **신뢰성**을 **객관적**으로 **확인**할 수 있도록 **검증절차·항목·방법**을 **마련**하고 제품·서비스의 검증 지원

< 예시 : 신뢰성 검증 항목 및 방법>

검증항목 : 안전성(안전기능 및 동작보장)	검증방법
■ **입력데이터의 검증 및 모니터링** ↳ △입력 **데이터 다중화** 여부, △**이상 데이터 학습 배제** 방식 적절성 등	■ **무작위 검증**(Randomized Test) ↳ 훈련 데이터의 확률적 일반성을 확인하기 위해 **임의로 조합된 데이터**를 **활용**
■ **모델의 오판 시 정상 동작** ↳ △사고 **시나리오별 동작 모드** 사전 정의 여부 △**위험 회피** 여부 등	■ **경계 조건 검증**(Corner Case Test) ↳ 발생빈도는 낮으나 결과가 위험할 수 있는 **시나리오**를 **설계**하여 **결과 위험** 여부 **확인**

○ 인공지능의 적용 목적, 활용 방식 및 파급력(심각도, 발생빈도 등)을 기반으로 **제품·서비스별 신뢰성 요구 수준** 및 **검증방안** 제시

< 예시 : 챗봇(저위험군)과 자율주행차(고위험군)의 안전성 요구 수준(※검토항목 √ 표시) >

안전성 속성 검증 항목	챗봇	자율주행차
■ **입력 데이터의 다중화**	-	√
■ **입력 데이터 비교 메커니즘**	-	√
■ **이상 데이터 학습 배제**	√	√
■ **정상 시나리오와 실시간 비교**	-	√
■ **상정 외 시나리오에 대한 실시간 학습 및 위험 회피**	-	√

□ (❸신뢰성 인증) **검증체계**를 **통과**(기술적·윤리적 요구사항 충족)한 인공지능 제품·서비스에 대해 **'신뢰성 인증'** 추진

○ **초기 시장**이고 제품·서비스별 **특수성**이 크며, **민간 자정 노력***이 진행 중임을 고려해 협·단체 중심의 **민간 자율인증**을 **우선 추진**

* '카카오' 알고리즘 헌장('18.1.), '네이버' 인공지능 윤리준칙 발표('21.2.), '삼성전자' 인공지능 윤리 제정 착수('20.12.), 'NC소프트' 윤리 제정 착수('21.3.)

- 제조社, 서비스社 등 **이해관계자**가 **참여**하는 **협의체**를 **구성**하여 **세부 인증요건***을 **마련**하고, 기술발전 양상 등을 고려해 지속 개선

* △'개발 가이드북' 준수한 개발 여부, △'검증체계'를 통한 검증 통과 여부 등

- **활용도**가 **높으면서 윤리 이슈**가 **예상**되는 **분야**(인공지능 스피커, 챗봇 등)를 중심으로 **시범 도입·운영**한 후 他분야로 점진적 확대

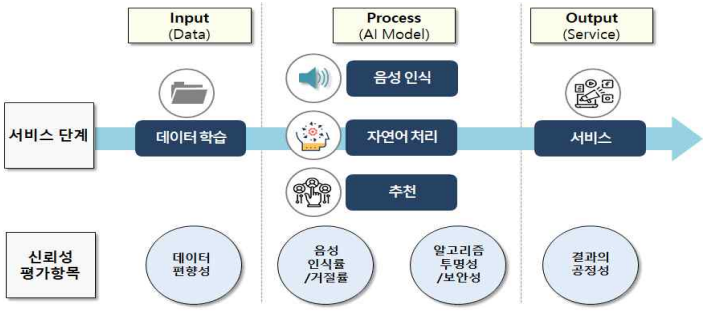
< 예시 : 인공지능 스피커 인증 항목 >

■ **서비스 구현 단계**를 **세분화**하여, 단계별 신뢰성 평가항목 도출

❶(입력) **학습 → 데이터 편향성**

❷(처리) **음성인식, 자연어처리**
↳알고리즘 **투명성**, **보안성** 등

❸(출력) **결과 → 공정성**



○ 다만, **생명·신체 등**에 **미칠 영향**이 **클 것**으로 **예상**되는 **분야**(의료, 교통 등)의 안전 관리를 위한 **정부 인증제 도입 검토** 추진

- 글로벌 정책 동향*, **국내 산업 발전 수준** 및 **위험도**가 큰 **인공지능**에 대한 **규제정책** 등과 **연계**하여 종합적으로 면밀한 검토 필요

* <EU 인공지능 법안> 고위험 인공지능에 대해 법적 요구사항 준수 여부 확인을 위한 적합성평가를 진행하고 통과 시 인증(CE 마크)하는 절차 제시

□ **(민간 자율 신뢰성 공시) 인증 제품**에 대한 **가치 제고**와 **인증 활성화 분위기 조성**을 위해 협·단체 주도로 **'신뢰성 공시'** 추진

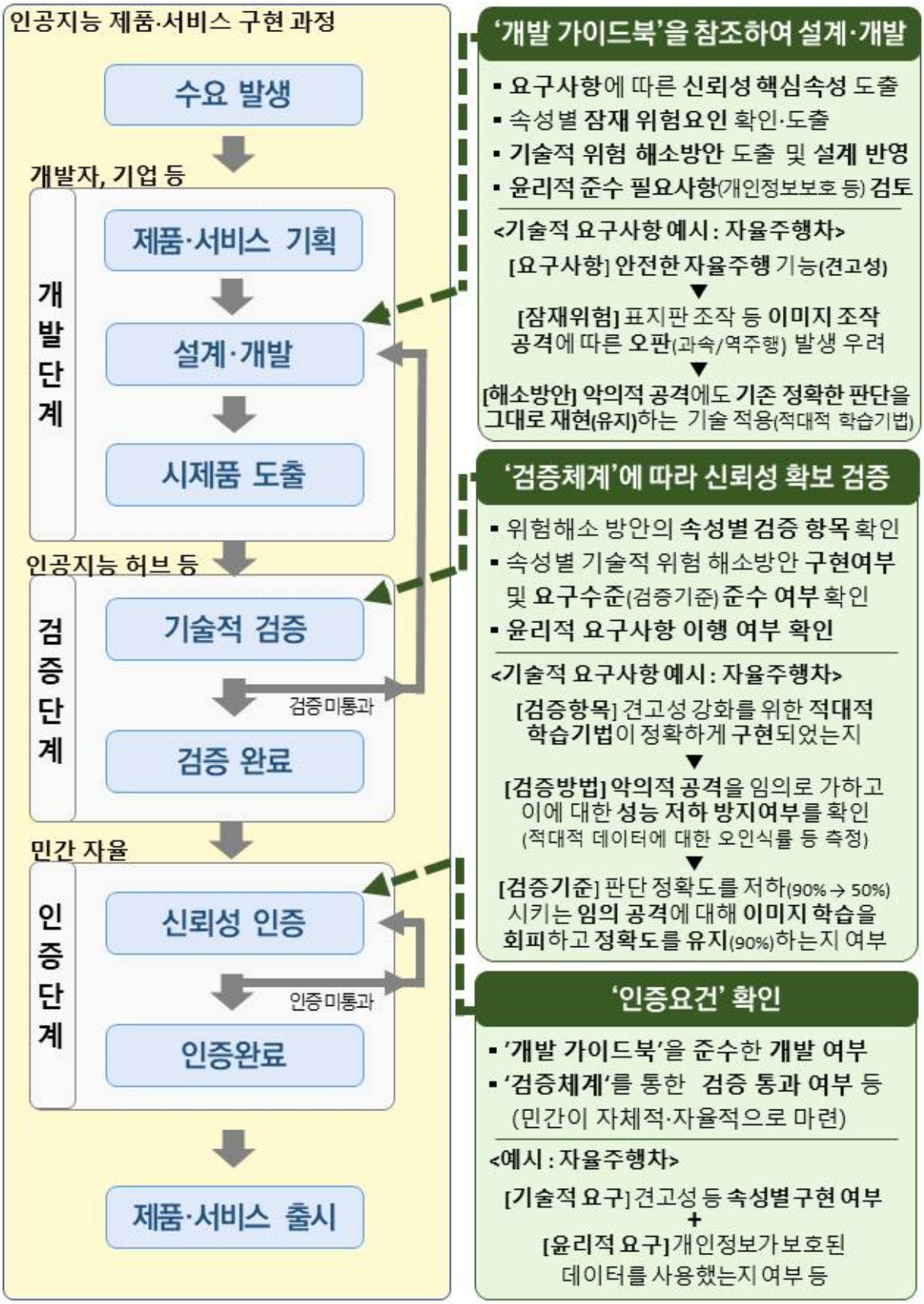
○ 인공지능 개발기업이 **자사의 신뢰성 관련 중요 정보***를 자발적으로 **온라인**에 **공시**하여 **기술력 공개** 및 **이용자 선택권 보장**을 지원

* 신뢰성 분야 △투자/활동, △인증, △인력 보유, △교육이수 현황 등

※ <사례 : 카카오> 자사 알고리즘 윤리헌장의 제·개정에 대한 정보를 온라인으로 공개

○ **협·단체**는 소속기업의 **공시제 참여**를 **유도·독려**하고 참여기업에 대해 **인증서 발급, 제품 홍보** 등 다양한 **인센티브 지원**

참고4 인공지능 개발 과정별 신뢰 확보 기준 적용 체계도



2. 민간 신뢰성 확보 지원

□ **(개발·검증 플랫폼 운영)** 기술·재정적 어려움을 겪는 **중소기업** 등의 **체계적 신뢰 확보 지원**을 위해 **One-Stop 지원체계** 구축

○ **(개발)** 제품·서비스의 **잠재 위험 분석**을 토대로 **신뢰성 속성별 검토사항, 개발방법(설계)론**을 **제시**해주는 **컨설팅·멘토링*** 추진

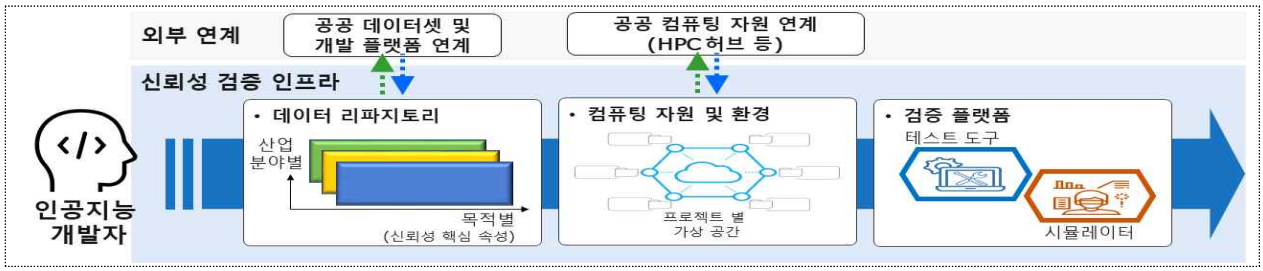
* 개발부터 시스템 검증까지 체계적·종합적 지원하는 '신뢰구현 바우처' 사업 추진 검토

○ **(검증)** 온라인에서 **검증체계**에 따른 **신뢰 속성별 수준 분석, 실(實)환경 테스트** 등이 가능한 클라우드 **검증 플랫폼** 구축·운영

- 기 구축·운영 중인 **인공지능 허브*** 등 공공 플랫폼과 **연계**하여 **'데이터 확보 → 알고리즘 학습(구현) → 검증'**을 **통합 제공**

* '학습용 데이터'(1,300종, ~'25년)와 '고성능 컴퓨팅'(상시 30.6PF 제공, '21년) 등 인공지능 개발에 필요한 인프라를 통합 제공하는 플랫폼('18~'20년간 29.4만명 접속)

< '데이터 확보 → 알고리즘 학습(컴퓨팅) → 검증' One-Stop 지원 플랫폼 구성안 >



○ **(노하우 공유)** 개발·검증 및 활용과정에서의 **신뢰성 구현 우수·모범 사례, 방법론** 등을 **산업 분야별**로 **수집·정리**하여 **공유**

□ **(민간 자율 인증 활성화)** 민간의 자율적 참여 확대 등 **산업 전반**의 **인공지능 신뢰 확보 분위기 조성**을 위해 **제도적 지원** 추진

○ **인증 제품**에 대한 **공공기관 우선 구매** 대상 품목 지정, **수의계약 지원, 조달 등록, 공공부문 사업 참여 요건** 반영 등 검토

○ 해외진출 지원 사업에 우선 포함, 국내·외 인공지능 박람회 등에 인증제품 출품 지원, 다양한 채널을 통한 홍보 등 추진

3. 인공지능 신뢰성 원천기술 개발

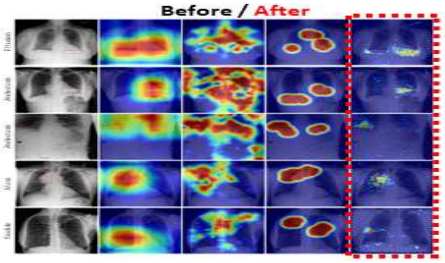
□ **(설명가능성)** 인공지능이 스스로 **판단**(의사결정)한 **기준·과정** 등을 **사람**이 **이해 가능한 방식**으로 **제시**(**설명**)할 수 있는 기술 개발

○ 인공지능 모델의 **인식 결과**에 대한 **설명가능성** 및 **예측 모델**에 대한 **판단 근거** 제공을 위한 **연구 수행** ※ [참고5] 참조

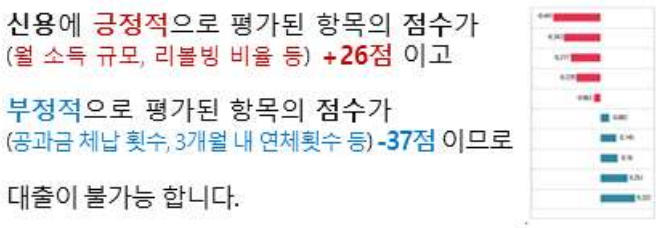
- 현재, 연구개발 결과를 **응용분야**(의료 영상진단, 금융지수 예측 등)에 **시범 적용**하여 **설명 가능한 인공지능 기술 적용·검증** 추진

< 참고 : 의료, 금융 분야 설명 가능한 인공지능 기술 적용 사례 >

■ **질병 진단**의 **판단 근거**를 **시각화**



■ **신용평가 결과**를 **시각·언어적**으로 표현



○ 플러그앤플레이*(Plug&Play) 방식으로 **기존 시스템에 설명가능성**을 **기능적**으로 **추가**할 수 있는 **범용 설명가능 기술개발** 신규 추진

* 새로운 기능을 사용자가 별다른 조치 없이 사용 가능하도록 지원하는 기술방식

- 단순 판단근거 제시를 넘어 보고서 생성, 대화 방식 등 **다양한 설명모드를 통한 사용자 맞춤형 설명가능성 제공 연구** 병행

※ (적용례 : 인공지능 손해사정사) 사고의 원인을 다각도로 분석하여 과실 비율을 결정하고 명확한 판단(판정) 근거가 포함된 사고처리 보고서 제공

□ **(공정성) 데이터·알고리즘**의 **편향 여부**를 **판단**하고, 인공지능이 **변화하는 가치**와 **규칙**을 **준수**할 수 있도록 하는 기술 개발

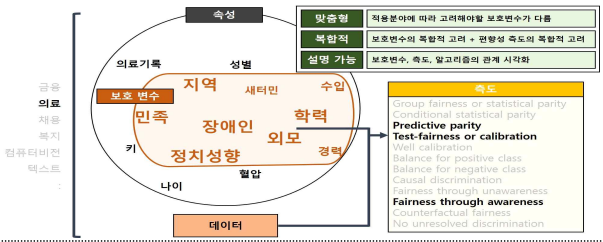
○ **분야별**(채용, 의료, 신용평가) **편향성 측도***를 기준으로 **학습데이터** 및 **알고리즘**의 **편향성 진단·제거 기술개발 중** ※ [참고5] 참조

* 편향성에 영향을 미치는 변수를 수학적으로 측정하기 위해 정의한 기준

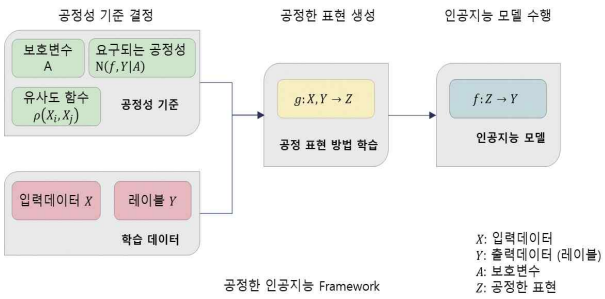
< 참고 : 공정한 인공지능 연구 추진 현황(KAIST) >

< 편향성 변수 추출 및 측도 개발(의료분야) >

■ 의료분야의 편향성 검증을 위해 지역, 학력, 외모 등을 측도로 정의한 사례



< 공정한 인공지능 프레임워크 개념도 >



○ 인공지능 윤리기준 등 관련 규정으로부터 **편향 가능성**을 **진단**하고 **편향성**을 **스스로 개선**(재학습, 알고리즘 수정 등)하는 **기술 신규 추진**

※ (적용례 : 인공지능 채용시스템) 매년 변경되는 채용 기준에 따라 스스로 평가 알고리즘(데이터)의 편향성을 진단·제거하고 새로운 기준에 맞춰 개선

□ **(견고성) 인공지능 알고리즘**에 대한 **적대적 공격**에도 유연하게 대응하여 **기능·성능**의 **변화 없이 안전**하게 **동작**하는 기술 개발

○ **인공지능**을 교란할 수 있는 **공격**의 **사례**를 **분석**하여 **기만공격**을 **사전**에 **감지**하고 **방어**할 수 있도록 하는 기술 개발 ※ [참고5] 참조

○ 기계학습(Machine Learning)을 통해 만들어진 **인공지능**의 보안상 **역기능·취약점**을 **자동 탐지**할 수 있는 기술 개발

- 음성, 영상 등 복합 정보를 이용한 **다양한 외부 공격에 선제적으로 대응**(방어·역공격 등)**하기 위한 기술 신규 추진** 병행

※ (적용례 : 인공지능 로봇) 다양한 정보(이미지, 영상, 음성, 위치 등)의 복합적 활용으로 발생 가능한 위해(危害) 유발 오인식·오작동의 예방·대응방안 마련

< 참고 : 인공지능에 대한 이미지 교란 공격 예시 >

■ 특정 사진을 **57.7%**의 **확률**로 **'판다'로 분류**하는 인공지능 시스템에 일부가 잡음(Noise)으로 **오염된 동일한 사진**을 **입력**했을 경우 **99.3%**의 **확률로 '긴팔원숭이'**로 **분류**하는 사례

참고5 현행 과제와 신규 연구과제 비교

□ 설명 가능한 인공지능 기술 개발

구분		현행 과제	신규 과제(안)
과제명		의사결정 이유를 설명 할 수 있는 인간 수준의 학습·추론 프레임워크	플러그앤플레이(Plug&Play) 방식의 설명 가능성 제공 기술 개발
기간/예산		'17~'21년(총 5년) / 총 187억원	'22~'26년(총 5년) / 총 450억원*
차이점	적용 단계	**모델 개발 시 설명가능성**을 **고려**하여 **설계**	설명가능성에 대한 고려 없이 **개발된 모델에 설명가능성 제공**
	설명 방식	**시각**적·**언어**적 설명	전문가용 **보고서 생성**, **대화** 등
	설명 수준	**개발자 수준**의 설명가능성 제공	**사용자 맞춤형** 설명가능성 제공

* 예타 통과 금액 기준

□ 공정한 인공지능 기술 개발

구분		현행 과제	신규 과제(안)
과제명		인공지능 모델과 학습데이터의 편향성 분석-탐지-완화·제거 지원 프레임워크 개발	적은 학습비용으로 변화하는 정책을 유연하게 준수하는 인공지능 기술개발
기간/예산		'19~'22년(총 4년) / 총 50억원	'22~'26년(총 5년) / 총 200억원*
차이점	적용 단계	**적용 분야별**로 **편향성 기준**을 **자체적**으로 **사전 정의**	**관련 규정 등**으로부터 **편향 요인을 스스로 식별**하여 **잠재적 편향 가능성 진단**
	편향성 제거방법	기준에 따라 **편향성**이 **제거·완화**된 **데이터 전체**를 **재학습**	부분데이터 재학습, 알고리즘 부분 수정 등 **최적의 방안**을 **스스로 판단·적용**

* 예타 통과 금액 기준

□ 견고한 인공지능 기술 개발

구분		현행 과제		신규 과제(안)
과제명		기만공격에 의한 인공지능 역기능 방지 기술 개발	기계학습 모델 보안 역기능 취약점 자동 탐지 및 방어기술 개발	복합지능 정보를 이용한 인공지능 공격에도 견고한 인공지능 기술 개발
기간/예산		'18~'20년(총 3년) / 총 13.65억원	'20~'27년(총 3년) / 총 46.5억원	'22~'26년(총 5년) / 기획 추진 중
차이점	연구 중점	인공지능 **기만 공격 사례 분석**을 통한 **기만 공격 사전감지** 기술 개발	인공지능 **역기능·취약점 자동탐지 시스템** 구축 기술 개발	**복합 정보**를 이용한 인공지능 공격에 대한 **선제적 대응** 기술 개발
	방어 대상	**단일 지능** 기반의 공격에 대한 방어		**복합 지능** 기반의 공격에 대한 방어

2 안전한 인공지능 활용을 위한 기반 마련

◇ 믿고 안전하게 인공지능을 **활용**할 수 있는 기반 조성을 위해 **고위험 인공지능 규제, 인공지능 영향평가 추진**

1. 학습용 데이터 신뢰성 제고

□ **(신뢰성 확보기준) 민·관**이 **학습용 데이터 구축**과정에서 **공통적**으로 **준수**해야 할 **표준 공정기준**(체계)을 **마련**하고 **활용 확산** 추진

○ **(표준 구축공정)** 데이터 유형별로 설계(기획)부터 구축(수집·정제·가공)까지 全주기 **포트폴리오** 및 **구축 가이드라인** 개발

○ **(공정별 요구사항)** 인공지능 활용 목적별*로 데이터 구축 **단계별 데이터 신뢰성** 확보를 위한 **상세 요구사항** 도출

* 시각지능(이미지·영상), 언어지능(텍스트·음성)을 통한 탐지, 추론 등

- 데이터 **신뢰성이 특히 민감한 일부 분야**(자율주행차, 금융 등)에 대해 특화된 분야별 신뢰성 요구사항 마련

○ **(검증지표)** 데이터 구축 **공정** 및 **결과물**의 신뢰성 요구사항 충족 여부를 확인하기 위한 **정량적 검증지표** 및 **측정방법** 마련

< 예시 : 인공지능 학습용 데이터 신뢰성 검증지표 >

구분	지표	주요 내용
데이터 구축 공정	준비성	법·제도(저작권, 개인정보보호 등) 검토, 데이터 구축계획(절차, 조직 등) 마련
	완전성	수집·정제·가공 계획의 체계성 및 준수 여부
	유용성	데이터 수요자(발주기관 등)의 요구사항 부합 여부 및 유연성
결과물 (데이터)	적합성	원시데이터의 통계적 다양성, 충분성, 사실성 및 포맷 준수 여부
	정확성	가공의 정확성, 정밀도 만족 여부 및 누락된 데이터 유무
	유효성	학습데이터로 훈련 시 분류·탐지·인식 등의 성능 수준 달성 여부

○ **(확산)** 주요 데이터 공급·수요기관을 중심으로 **협의체** 및 **자문 채널 운영, 안내서 보급** 등을 추진하여 통일성 있는 활용을 촉진

□ **(데이터댐 - 학습용데이터 신뢰성 강화)** '기획 → 수집 → 가공 → 개방 → 활용' 全과정에 **신뢰성 확보**를 위한 **고려사항 적용·운영**

< 인공지능 학습용 데이터 구축사업 개요 >

■ **「디지털뉴딜」**의 핵심과제로 **인공지능 학습에 필요한 질 높은 데이터**를 **'25년까지 누적 1,300종 구축** 및 **개방**하는 '데이터 댐' 프로젝트

■ '21.5월 현재 **8대 분야***별 **총 191종**의 **학습용 데이터**를 **구축**하여 '**인공지능 허브**'를 통해 단계적으로 **개방 중**이며, 스타트업 등의 서비스·제품 개발에 **누적 5만건 활용**

* ①자연어, ②헬스케어, ③자율주행, ④농축수산, ⑤국토·환경, ⑥미디어, ⑦안전, ⑧지역특화

○ **(기획)** 인공지능이 **해결할 문제**와 문제 해결에 필요한 **학습용 데이터**의 수집·가공 방식, 규모 등을 **명확**하게 **정의**하고 **설계**

- **구축 全단계**(수집·가공·검수)에서 발생할 수 있는 **잠재적 신뢰성 저하 요인**을 데이터별 특수성에 따라 **사전 검토·정의**

○ **(수집)** 원천 데이터의 **다양성, 충분성, 사실성, 공정성**을 확보하고, **획득과정**에서 저작권, 개인정보보호 등 **법·제도 준수**

○ **(가공)** 데이터의 **유형, 학습 목적**에 적합한 **라벨링** 및 **작업방식**(크라우드소싱, 교차검증 등)을 **채택**하여, 가공 데이터의 **정확성** 확보

○ **(개방)** 구축된 데이터 **개방** 전, **통계적 다양성, 가공 정확성** 등을 정량적으로 **평가·검증**하여 **품질·신뢰성 확보**

○ **(활용) 데이터 오류 신고 창구 운영**(온라인)을 통해 **데이터 활용 과정**에서 발견하는 **오류**를 **확인**하고 **지속 보완***하여 품질 유지

* 신고사항은 △데이터 구축기관에 보완조치, △차년도 품질관리 계획 반영하여 고도화

- 구축 데이터에 **사후 유지보수 기간**(1년)을 설정하여, **오류**가 **발생** 시 **구축기관**의 **책임** 하에 **체계적·종합적 보완**을 실시

- 데이터 오류는 아니나, 인공지능 **성능 향상** 등을 위해 **필요 시 추가·연계 과제***를 **기획·추진**하여 **데이터 고도화**

* (예) 데이터 추가(일반도로 → 경사진 도로 등), 가공 대상 추가(자동차 → 보행자, 장애물 등), 가공 방식 추가(바운딩박스 → 세그멘테이션 등)

2. 고위험 인공지능에 대한 신뢰 확보

□ **(고위험 인공지능 범주 설정)** 이용자 보호 등을 위해 **사람**에게 **잠재적 위험**을 **미칠 수 있는 인공지능**에 대해 **별도 구분·분류** 추진

○ 글로벌 동향 검토, 정책연구, 각계 의견수렴 등을 통해 범주 설정

※ △EU는 분야·방법에 따른 고위험 인공지능 범주를 제시('20.2)하고, 「인공지능 규제안」에서 생체인식 시스템 등 8개를 규정('20.4), △미국 '위험'기반 사후규제 원칙 제시('20.1.)

□ **(신뢰 확보 제도)** **고위험 인공지능**이 **활용**되는 **서비스**에서 **공급자**(기업 등)↔**이용자** 간 **상호 신뢰**를 가능토록 하는 **제도 도입** 추진

< 인공지능 서비스 단계에 따른 주체별 의무/권리(안) >

○ 고위험 인공지능을 활용한 서비스인 점에 대해 상호 동일한 인식을 할 수 있도록 **서비스 제공前 인공지능 활용여부**의 **'고지'**를 **의무화**

○ 고지 이후 **해당 서비스**에 대한 **'거부'**, 인공지능의 **판단 근거**에 대한 **'설명'** 및 이에 대한 **'이의제기' 등**에 대해 **중장기 검토**

- 글로벌 입법·제도화 동향, 산업적 파급력, 사회적 합의·수용성, 기술적 실현 가능성 등에 대해 **다각적**으로 **검토 필요**

□ **(실행방안)** 산업계 **의견 수렴** 등 **사회적 합의**를 거쳐 **기존 법안**(「지능정보화기본법」, 「개인정보보호법」) **개정** 또는 **신규 법안 제정** 추진

[예시 : 고위험 인공지능을 활용한 수술에서의 이용자 권리 및 사업자 의무]

참고6 국내·외 구속력 있는 신뢰 확보 제도 관련 사례

1 EU 「인공지능 규제안」(AI regulation, '21.4.) 中 고위험 인공지능

○ 시민들의 **안전**이나 **기본권**에 **부정적인 영향**을 끼치는 **인공지능 시스템**으로, 동 법률에서 목록으로 제시되는 시스템 등

[구체적 대상 목록]

① **생체인식 시스템**
② 교통, 수도, 가스, 난방, 전기 등 **중요 기반시설**의 **관리** 및 **운용**에 활용
③ **교육/직업훈련**에서 시험 채점과 같이 액세스를 결정하거나 **평가, 배치 결정** 등에 **활용**
④ **고용, 근로자 관리** 및 **자영업**에서 **채용·승진** 등의 **이력서 스크리닝, 평가** 등에 활용
⑤ **응급 서비스, 대출 신용평가** 등 필수 공공/민간 서비스에 활용
⑥ **수사/기소과정**에서 **증거**의 **신뢰성 평가** 등 기본권을 간섭할 수 있는 법 집행에 활용
⑦ **이민, 망명** 및 **국경통제 관리**에서의 **문서**의 **진위 확인**, **위험 평가** 등
⑧ **사법 과정**에서의 **사실**과 **법의 조사** 및 **해석, 법 적용 지원** 등

※ 「인공지능백서」('2021)에서 고위험 인공지능을 △상당한 위험 발생이 예상되는 '분야'에서 사용되고, △해당 분야에서 상당한 위험이 발생할 가능성이 높은 '방법'으로 사용되는 경우로 제시

2 이용자 신뢰 확보를 위한 제도

○ **(EU : GDPR) 자동화된 의사결정**에 대한 사업자의 △**사전 고지** 및 **이용자**의 △**이용 거부**, △결과에 대한 **설명 요구**, △결과에 대한 **이의제기 권리 보장**을 규정('16.5. 제정, '18.5. 시행)

■ **(제22조)** ①개인정보주체는 프로파일링 등, 본인에 관한 법적 효력을 초래하거나 이와 유사하게 본인에게 중대한 영향을 미치는 자동화된 처리에만 의존하는 결정의 적용을 받지 않을 권리를 가진다. ②결정이 다음 각 호에 해당하는 경우에는 제1항이 적용되지 않는다.
(a)개인정보주체와 개인정보처리자 간의 계약을 체결 또는 이행하는데 필요한 경우
(c) 개인정보주체의 명백한 동의에 근거하는 경우

○ **(국내 : 개인정보보호법)** 정보주체에게 법적 효력 또는 중대한 영향을 미치는 **자동화된 의사결정**에 대한 △**사전고지**, △**거부**, △**이의제기**, △**설명요구**를 보장('21.1.6. 입법예고)

■ **(제37조의2)** ①정보주체는 (생략) 자동화된 개인정보 처리에만 의존하여 특정 정보주체에게 개별적으로 법적 효력 또는 생명·신체·정신·재산에 중대한 영향을 미치는 의사결정을 행한 개인정보처리자에 대하여 그 **거부**, **이의제기**, **설명** 등을 **요구**할 수 있다. (생략) ③개인정보처리자는 제1항에 따른 자동화 의사결정의 기준과 절차를 (생략) 정보주체가 **사전에 쉽게 인식**할 수 있도록 알리는 등 필요한 조치를 하여야 한다.

3. 인공지능 영향평가 추진

□ **(인공지능 영향평가)** 인공지능이 **국민생활 전반**에 **미치는 영향**을 **체계적**으로 **분석**하여 **대응**하기 위해 **사회적 영향평가 도입*** 추진

* 「지능정보화기본법」 제56조 : 국민의 생활에 파급력이 큰 지능정보서비스 등의 활용과 확산이 사회·경제·문화 및 국민의 일상생활 등에 미치는 영향에 대한 영향평가를 할 수 있다.

※ 개인정보 보호에 미치는 영향에 관하여는 개인정보 영향평가를 통하여 평가(「개인정보보호법」 제33조)

○ 인공지능에 따른 사회 변화, 영향력 등에 대해 **연속적·종합적 관점**에서 **조사·분석**하기 위해 **정례적·다면적**으로 **실시**

- 안전성, 투명성 등 신뢰성 요소를 토대로 평가하여 **인공지능**의 **영향력**을 **종합 분석**하고 **기술·관리적 조치방안**에 대해 **제시·권고***

* (예) 부정적 영향력의 회복가능성, 지속성을 기준으로 4등급(극소-소-중-대)으로 구분하고, 각 등급에 따라 해당 서비스의 투명성, 설명가능성, 모니터링 개선 요구를 차등화

○ 파급력이 큰 **공공 서비스**(지능형 교통시스템, 자동화 행정)에서 **시범 적용·운영**한 후 성과 평가 등을 거쳐 민간서비스로 단계적 확장

< 주요국 인공지능 영향평가 사례 >

구분	캐나다	EU
도입	'19.4 시행	'20.7 제안
대상	**정부**의 자동화된 의사결정 시스템	**민간 기업**의 인공지능 시스템
평가 항목	■ 자동화된 의사결정이 미치는 영향에 따라 구분된 4단계별 **요구사항**(판단 기준 설명, 사전고지 등) **충족 여부**	■ **사람의 기본권에 미치는 영향** (차별, 개인 정보보호 등) ■ **인공지능 시스템의 요건 충족 여부** (안전성, 견고성, 투명성 등)
기타	■ 자동화된 행정결정에 대한 도입 전 영향평가의 의무 실시 필요	-

□ **(산업·기술 전반 조사)** 효과적인 영향평가의 추진이 가능하도록 **국내·외 인공지능 산업·기술**의 **동향·실태** 등에 대한 **조사 실시**

○ 국내 산업·서비스별 **인공지능 도입수준, 기술적 특성** 및 **주(主) 이용자 현황** 등 사회 전반의 인공지능 활용에 대해 면밀히 조사

○ **국외 영향평가 추진 동향**(신뢰성 제도 등)에 대한 정책 분석·연구

참고7 국내·외 영향평가 제도

□ 국내 : '사회적 영향평가', '기술영향평가'

구분	사회적 영향평가	기술 영향평가
근거	「지능정보화기본법」 제56조	「과학기술기본법」 제14조
대상	국민의 생활에 파급력이 큰 **지능정보서비스** 등	**미래**의 **신기술** 및 **기술적·경제적·사회적 영향**과 **파급효과** 등이 **큰 기술**
평가 항목	■ **사회·경제·문화** 및 **국민의 일상생활** 등에 **미치는 영향**과 관련된 내용 ①지능정보서비스 등의 안전성 및 신뢰성 ②정보격차 해소, 사생활 보호, 지능정보 사회윤리 등 정보문화에 미치는 영향 ③고용·노동, 공정거래, 산업 구조, 이용자 권익 등 사회·경제에 미치는 영향 ④정보보호에 미치는 영향 ⑤그 밖에 지능정보서비스 등이 사회·경제·문화 및 국민의 일상생활에 미치는 영향	■ 새로운 과학기술의 발전이 경제사회문화 윤리환경 등에 미치는 영향과 관련된 내용 ①해당 기술이 국민생활의 편익증진 및 관련 산업의 발전에 미치는 영향 ②새로운 과학기술이 경제·사회·문화·윤리 및 환경에 미치는 영향 ③해당 기술이 부작용을 초래할 가능성이 있는 경우 이를 방지할 수 있는 방안

□ 국외 : 캐나다, EU

구분	캐나다('19.4. 시행)	EU('20.7. 제안)
근거	자동화된 의사결정에 대한 지침 (Directive on Automated Decision-Making)	신뢰가능한 인공지능을 위한 자체 평가 목록 (The Assessment List for Trustworthy Artificial Intelligence for self assessment)
대상	**정부**의 **자동화된 의사결정 시스템**	**기업**의 **인공지능 시스템**
평가 항목	■ 자동화된 의사결정이 미치는 영향에 따라 **4단계**로 **구분**하여 **요건 적용** ◆알고리즘 영향평가 ①자동화된 의사결정 시스템 생산 전 알고리즘 영향평가를 완료 ②알고리즘 영향평가 결과에 따라 부록C에 규정된 관련 요건을 적용 ③자동화된 의사결정 시스템의 기능 또는 범위가 변경될 시 알고리즘 영향평가를 갱신 ④알고리즘 영향평가의 최종 결과를 일반 접근이 가능한 형식으로 공개 ◆투명성 관련 ①(사전고지) 부록C에 따라 자동화된 의사결정 시스템에 의해 전체 또는 부분적으로 수행된다는 내용을 고지 ②(사후설명) 부록C에 따라 결정이 내려진 방법과 이유에 대해 영향을 받는 개인에게 이해 가능하도록 설명 ③소프트웨어 구성 요소에 대한 접근 권한 ④사용자 정의 소스 코드 공개	■ 인공지능의 부정적 영향 최소화 등 **신뢰가능한 인공지능 구현** 관련 내용 ◆인간의 기본적 권리에 미치는 영향 ①부정적 차별 가능성 ②아동의 보호와 권리 보장 여부 ③GDPR에 따른 개인 데이터 보호 여부 ④표현·정보의 자유, 집회·연대의 자유 보장 여부 ◆인공지능 시스템의 7가지 요건 준수 여부 ①인간에 의한 관리감독 ②기술적 견고성 및 안전성 영향 ③프라이버시에 미치는 영향 및 데이터 거버넌스 구축 ④투명성 ⑤다양성, 차별금지 및 공정성 ⑥사회 및 환경에 미치는 영향 ⑦책임성

4. 사회 전반 신뢰 강화 제도 개선

□ 사회 전반에서 활용되는 **인공지능**에 대한 **신뢰 확보, 이용자**의 **생명·신체 보호** 등을 위한 **인공지능 관련 법·제도 정비*** 추진

* 「인공지능 법·제도·규제 정비 로드맵」('20.12., 관계부처)을 통해 발굴된 과제 4건

☞ **민간**의 **자율성**을 **저해하지 않는 선**에서, **글로벌 입법 동향, 인공지능**의 사회·산업적 **파급력** 및 **기술 발전 수준** 등을 다각도로 면밀히 고려하여 **제도 개선**의 **범위**와 **대상**을 **명확히** 하고 **불확실성**을 **제거**

① 업계 자율적 알고리즘 관리·감독 환경 조성(과기정통부, '21.下~)

○ **알고리즘**의 오류 등을 **평가·검증**할 수 있는 **체계**가 **부재**한 것을 **보완**하기 위해 **표준 관리·감독체계**에 대한 **가이드라인*** **제정**

* 사업자가 자율적·지속적으로 신뢰성·투명성에 대해 추적·평가·관리할 수 있도록 지원하는 절차, 모델, 검증방법 등을 포함한 현장 실행지침

② 플랫폼 알고리즘 공정성·투명성 확보(공정위·방통위·과기정통부, '21.上)

○ 플랫폼 사업자의 **인위적**인 **알고리즘 조작**에 따른 **피해 방지**를 위해 알고리즘의 **공정성 검증기준·절차**의 제도화 등 보완책 마련

③ 영업비밀 보장을 위한 알고리즘 공개 기준 마련(과기정통부·공정위, '21.下)

○ **민간**의 **자율적**인 **알고리즘 투명성 확보 노력 지원을** 위해 영업비밀을 침해하지 않는 범위 내 **알고리즘 공개기준** 등*을 **검토·마련**

* 영업비밀의 판단 기준, 알고리즘 공개·설명의 범위·방법 등 세부사항 포함

④ 고위험 분야 기술기준 마련(과기정통부, '22)

○ **생명·신체 등**에 **밀접한 분야**(의료 등 고위험분야)에서 사업자가 준수해야 할 **인공지능**의 **안전성·신뢰성** 등에 관한 **기술기준** 제시

- 인공지능에 대한 국민 안전·신뢰 확보를 위해 현재 산업성숙도 등을 고려해 유보 중인 기술기준(「지능정보화기본법」 제21조*)에 대한 제정 추진

* 국민의 생명·신체안전 등에 밀접한 지능정보기술에 관련된 사업자는 과기정통부장관이 정하여 고시하는 기준에 적합하도록 지능정보기술을 개발·관리·활용하여야 한다.

3 사회 전반 건전한 인공지능 의식 확산

◇ 건전한 인공지능 활용 의식 확산과 분위기 조성을 위해 **인공지능 윤리 교육 강화, 체크리스트 보급** 등 추진

1. 인공지능 윤리교육 강화

□ **(윤리교육 총론) 국가 인공지능 윤리교육**의 **방향**을 **제시**하는 총론 개발

○ 인공지능이 사회에 미치는 영향, 인간-인공지능간 상호작용 등 **사회·인문학적 관점**과 **윤리기준**의 **사회실천***을 인식할 수 있는 내용 반영

* '인간성을 위한 인공지능'(AI for Humanity), 3대원칙과 10대 요건 등 윤리기준 주요내용을 일상에서 실천할 수 있도록 내용 구성하고 시나리오를 제시

○ **인공지능 윤리교육 목표**(연령, 개발자·이용자별 등), **핵심 교육과정 기준안, 주체(직업)별 필수역량** 등 교육에 필요한 세부사항 **개발·제시**

□ **(교육과정 개발)** 연구·개발자, 일반시민, 초중고생 등 **주체별·단계별 특성**을 **고려**한 **맞춤형 인공지능 윤리 교육과정 개발**

○ **(일반시민) 全국민 디지털 역량 강화 프로그램**과 **연계**하여 디지털 역량 수준*에 맞는 **인공지능 윤리교육 커리큘럼**을 **개발**

* 디지털 역량 수준 진단을 위한 측정 방법을 개발 중이며, 역량 수준별 맞춤형 교육 프로그램 설계 및 제공을 추진 중(디지털 포용 추진계획, '20.6)

○ **(연구·개발자)** 주요 산·학·연과 함께 **업무, 직장·사회생활** 등과 「인공지능 윤리기준」을 연계*한 **학습 및 훈련과정 개발**

* (예) 윤리기준을 기업 조직 가치와 연계하여 분석, 체크리스트의 핵심 준수사항에 대한 이해, 인공지능 활용 목적 및 맥락별 실천(적용) 단계의 고려사항 토의 등

- 인공지능대학원, SW중심대학 등의 인공지능 기본 역량교육에서 활용

< 예시 : 연구자·개발자 대상 인공지능 윤리 교육 목표 및 내용 >

[교육목표]

■ **인공지능기반**의 **판단**이 **인간**의 **존엄성을 침해**하는지를 **판단**하고, **제어**할 수 있는 **역량 훈련**

■ 사회문화적 가치를 반영한 **윤리성**, **이해 상충 시 판단 능력**, **책임성**, **사고**와 **소통 역량 함양**

[예시]

①**인공지능 시스템**에 대한 **이해관계자 식별 및 가치 도출, 윤리 매트릭스 작성**

②공동체 가치 구현을 위한 **인공지능 시스템**의 새로운 **목표 설정** 및 **훈련용 데이터셋 식별**

③새로 설정한 목표 및 **이해관계자 가치**를 **반영**한 인공지능 **설계**

④인공지능 시스템으로 인해 발생할 수 있는 **사회적 영향 탐구** 및 **논의**

○ **(초중고생)** 정보 관련 교과* 내 **인공지능**의 **사회적 영향**에 대한 **윤리 교육 강화**를 위한 **내용요소 개발** 및 **반영**('25~)

* 초등 '실과' 및 중등 '정보'(현행 2015 개정 교육과정 기준)

- **'창의적체험활동', '자유학기제'** 등을 활용하여 **학교 자율적**으로 **인공지능 윤리** 등 디지털 리터러시 관련 **교육**을 **실시** 할 수 있도록 **안내**

- **인공지능 윤리 교육**을 위한 **과목 개설** 및 **교과서 개발**은 **시·도 교육청**(학교)별 **자율**로 추진

2. 주체별 체크리스트 마련

□ **(체크리스트)** 사람이 중심이 되는 「인공지능 윤리기준」('20.12.)'의 구체적 행위지침으로써 **주체별*** **윤리 체크리스트 개발**('21~'22.)

* (예) 인공지능 기술 연구·개발자('21.), 인공지능 기반 제품·서비스 제공자 및 이용자('22.)

○ 법·윤리·기술전문가, 시민 등 **다양한 사회 구성원**과 함께, **인공지능 개발·활용** 시 **자체적**으로 **점검**해야할 **핵심사항 도출** 추진

○ **기술발전 양상**을 **반영**하고, **他분야 체크리스트**와도 **정합성***을 **유지**(주기적인 기술·사회적 검증 실시)하여, 현장의 **실천 가능성 제고**

* 인공지능과 관련된 분야(개발, 활용, 개인정보 등)의 체크리스트는 「인공지능 윤리기준」을 기본 원칙으로 하여 통일성·체계성 있게 마련 추진

< 체크리스트 문항 예시(안) >

체크리스트 문항	윤리기준 요건
■ 인공지능 시스템이 **개인정보를 활용 또는 처리**하여 **훈련**하거나 **개발**되었는가?	사생활 보호, 데이터관리
■ 인공지능 개발 단계에서 사회에 존재하는 선호의 **다양성**, **이용자 간 수용능력 차이**를 **고려**하였는가?	다양성 존중, 공공성
■ 훈련데이터 출처 기록 등 감사를 위해 인공지능 **개발 과정**을 **추적할 수 있는 프레임워크**를 **구축**하였는가?	책임성, 투명성
■ 인공지능 시스템 **결과**를 **사용자에게 설명**할 수 있는가?	투명성

□ **(홍보·확산)** 주체별 인공지능 윤리 체크 리스트의 **확산**과 **사회 구성원**의 **적극적 실천·이행**을 위해 홍보·확산 방안 마련

○ 언제 어디서나 **체크리스트**를 **자율점검** 할 수 있고, 타 사례를 토대로 **보완사항** 등의 **피드백**을 **제공**하는 **온라인 점검도구*** 보급

* <사례 : EU> 개인이 평가목록을 간편하게 검증할 수 있는 도구를 웹사이트에 공개

○ 국내·외 실제 윤리 이슈를 바탕으로 **체크리스트**를 쉽게 설명한 **해설서 개발·보급** 등 주체별 맞춤형 확산 방안*을 검토 추진

* (예) 민간자율 인공지능 신뢰성 인증제도 항목에 윤리 체크리스트 활용 포함

3. 인공지능 윤리 정책 플랫폼 운영

□ **(인공지능 윤리 포럼) 학계·기업·시민단체·공공 등**이 참여하여 **인공지능 윤리**에 대해 깊이 있게 **토의**하는 **공론의 장 마련** 추진

□ **(의견수렴 플랫폼)** 인공지능 윤리에 대한 **개발자, 이용자** 등 **사회 구성원**의 **다양한 의견**을 **수렴**하는 **온라인 플랫폼 운영** 추진

○ 전문가와 대중이 자유롭게 의견을 내고 토론하는 多방향 소통의 창으로써, 윤리 전반에 대한 폭넓은 의견수렴의 장으로 운영

※ 시범적으로 인공지능 기술 연구·개발자 체크리스트('21)에 대한 공론화 추진

< 사례 : EU "The European AI Alliance" >

■ EU 집행위는 **인공지능 분야 온라인 시민 의견수렴 플랫폼**인 "The European AI Alliance(유럽 인공지능 연합)" 운영

■ 초창기에는 **'인공지능 고위 전문가 그룹(AI HLEG)' 활동에 대한 피드백 제공**에 집중했으나, 점차 **4천여명 이상의 다양한 이해관계자가 참여**하여 **인공지능 관련 논의를 주도**하고 **유럽위원회의 정책결정에 기여**하는 시민참여 플랫폼으로 성장

Ⅵ. 추진 일정

구분	'21년	'22년	'23년	'24년	'25년
신뢰 가능한 인공지능 구현 환경 조성					
[1-1] 인공지능 제품·서비스 신뢰 확보 체계 마련					
▪ 신뢰성 기반 '개발 가이드북' 개발·보급	속성/요구사항 도출	분야별 적용	분야별 고도화		
▪ 신뢰성 '검증체계' 개발	검증항목/수준 도출	수준별 체계 고도화	분야별 적용 및 고도화		
▪ 민간 자율 '신뢰성 인증제' 운영	인증체계 연구	시범 도입 및 실증		他분야 확산	
▪ 고위험 분야 정부 인증제 도입 검토		기획	시범운영	적용 확산 및 고도화	
▪ 민간 인공지능 신뢰성 공시	기획	시범 도입	시행/확산		
[1-2] 민간 신뢰성 확보 지원					
▪ 민간 신뢰성 확보 지원		검증 플랫폼 구축	시범운영 및 고도화	검증 지원	
		시범수행	컨설팅·멘토링 수행	기업 지원	
▪ 우선구매 품목지정 등 제도적 지원 검토	운영방안 검토/연구		제도 개선 추진	제도 운영	
[1-3] 인공지능 신뢰성 원천기술 개발					
▪ 설명가능한 인공지능 기술 개발	응용분야 적용·검증				
	예타	세부 기획	기술 개발		적용/실증 등
▪ 공정한 인공지능 기술 개발	편향성 측정 기술개발(의료, 채용 등)				
	예타	세부 기획	기술 개발		적용/실증 등
▪ 견고한 인공지능 기술개발	인공지능의 역기능·취약점 자동탐지 기술 개발(~'27)				
		과제기획	기술 개발		적용/실증 등
안전한 인공지능 활용을 위한 기반 마련					
[2-1] 학습용 데이터 신뢰성 제고					
▪ 데이터 공정별 요구기준·검증지표 마련		표준공정 등개발	확산 및 고도화		
▪ 학습용데이터 구축(데이터댐) 全과정 신뢰 확보	가이드라인	적용·운영 및 고도화			
[2-2] 고위험 인공지능에 대한 신뢰 확보					
▪ 고위험 인공지능 범주 설정 검토	제도화 연구(해외동향 등)		초안마련/의견수렴		수립/시행(제도화)
▪ 이용자 신뢰성 확보 제도 도입 검토	제도화 연구(해외동향 등)		초안마련/의견수렴		수립/시행(제도화)
[2-3] 인공지능 영향평가 추진					
▪ 인공지능 영향평가 방안 마련		지표/위험분류체계 개발	공공분야 시범운영/지표 등 고도화		민간분야 확대
▪ 산업·기술 전반 조사		제도분석	산업별 도입 기술 및 활용 수준 조사	국내외 신뢰성 제도 간 정합성 연구	
[2-4] 사회 전반 신뢰 강화 제도 개선					
▪ 알고리즘 표준관리·감독체계 가이드라인 마련		제도화 연구		초안마련/의견수렴	
▪ 플랫폼 알고리즘 관련 법규 정비	플랫폼공정화법 등 발의		하위법령 정비		
▪ 기업 알고리즘 공개 기준·가이드라인 마련		기술현황·해외동향 등 연구	가이드라인 연구		초안마련/의견수렴
▪ 인공지능 기술기준 고시 제정		기술현황·해외동향 등 연구	제도화 연구		초안마련/의견수렴
사회전반 인공지능 의식 확산					
[3-1] 인공지능 윤리교육 강화					
▪ 인공지능 윤리 총론 개발	개발/의견수렴				
▪ 주체별·단계별 맞춤형 교육과정 개발		연구자/개발자/대학용 개발		현장 적용 및 고도화	
		초중등 교육과정 개발 및 학교 적용 검토(교육부 협의)			교육과정 적용
[3-2] 주체별 체크리스트 마련					
▪ 주체별 윤리 체크리스트 개발·보급	개발/의견수렴		현장 시범적용 및 고도화 및 정책적 인센티브 검토		
[3-3] 인공지능 윤리 정책 플랫폼 운영					
▪ 인공지능 윤리 포럼 운영	기획/시범운영	윤리 전문가 그룹 발족 및 운영			
▪ 시민 의견 수렴 플랫폼 구축·운영		플랫폼 구축	운영 및 플랫폼 고도화		

Ⅶ. 추진 체계

◇ 관계부처·민간과 적극적으로 소통하며 체계적 추진