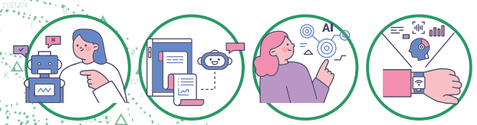
# 인공지능이 펼칠 혁신에 믿음을 더합니다.

TTA AI융합시험연구소

자세한 정보는
QR코드를 촬영하세요

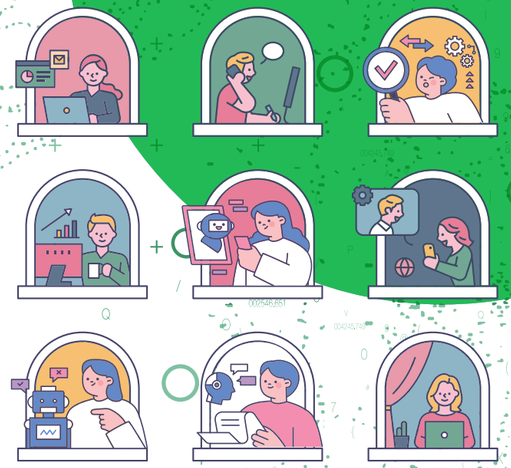

# TOWARD TRUSTWORTHY AI



TTA 한국정보통신기술협회
Telecommunications Technology Association

경기도 성남시 분당구 분당로 47(서현동 267-2)
TEL. 031)724-0001 trustworthyai@tta.or.kr

# 신뢰할 수 있는 인공지능 컨설팅



과학기술정보통신부 Ministry of Science and ICT
TTA 한국정보통신기술협회 Telecommunications Technology Association

# 컨설팅 정보

## 추진목적

- 인공지능 제품·서비스 개발 또는 운영 과정에서 발생 가능한 윤리적 이슈, 위험 발생 요소 등을 사전 진단하고 개선을 지원함으로써 신뢰성*이 보장된 인공지능 기술 확산

  * 인공지능 신뢰성(AI Trustworthiness) : 데이터 및 모델의 편향, 설명이 어려운 인공지능 기술이 내재한 위험과 한계를 해결하고, 인공지능 활용 과정에서 부작용을 방지하기 위해 준수해야 하는 가치 기준

## 컨설팅 대상

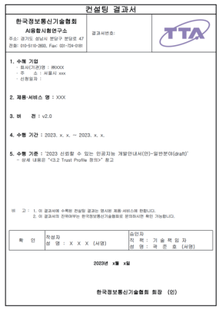
- 인공지능 제품·서비스를 제공을 목표로 하는 기업

## 컨설팅 내용

- 신뢰할 수 있는 인공지능 개발 안내서를 기반으로 컨설팅 대상 제품·서비스에 적합한 Trust Profile*을 도출하여 지속적인 신뢰성 확보 내재화를 지원

  * Trust Profile : 대상 제품·서비스가 적용해야 하는 요구사항과 검증항목

1. **신뢰성 교육진행** — 인공지능 신뢰성 개념 및 프레임워크 소개
2. **Trust Profile 확정** — 대상 제품·서비스의 신뢰성 고려 요소 분석
3. **Trust Profile 진단** — 진단 결과 안내
4. **개선 여부 확인** — 미준수 요구사항에 대한 개선방안 제안




"제품·서비스에 대한 분석 결과와 상세한 진단·개선사항이 기술된 컨설팅 결과보고서 제공"

"컨설팅 진행 과정에서 발생한 중요 정보에 대해 철저한 보안 유지"

# 인공지능 신뢰성 확보 현황 분석 예시

## 분석 대상 솔루션 정보

- **기업명** A 기업
- **솔루션명** B 솔루션 (AI 얼굴 인식 솔루션)
- **현황** 개발 완료 및 운영 중
- **서비스 형태** 클라우드 기반 AI Engine(SaaS)을 통해 얼굴 영상 분석 및 감지 결과 제공

## 인공지능 신뢰성 확보 현황 분석 요약

- **확인 대상**
  1. B 솔루션 Engine : 데이터 소스로부터 영상 데이터를 취득한 후 클라우드 플랫폼을 통해 가공하여 인공지능 모델 학습을 진행하고, Deploy된 인공지능 모델을 활용하여 실시간 영상 데이터 중 얼굴을 인식한 결과를 End-user에게 전달하는 SW기능
  2. B 솔루션 Service : B 솔루션 Service 부서에서 진행하는 고객 대응 및 피드백 전달, 인공지능 모델 판단 결과의 모니터링 및 검토

- **종합의견**

  B 솔루션은 '신뢰할 수 있는 인공지능 개발 안내서'에 명시된 요구사항 15개 중 9개 요구사항을 인공지능 생명주기 전 단계에 걸쳐 고려하여야 하는 것으로 분석되었습니다.
  아울러, 9개 요구사항의 충족 여부 검토 결과 B 솔루션은 Engine 및 Service 대부분 충족하고 있습니다.
  향후 B 솔루션의 지속적인 신뢰성 확보를 위해 사내 구성원과 함께 권고사항의 적용을 고려하시기 바랍니다.

- **Trust Profile 분석 결과** (준수율 : 77%)

| 구분 | 생명주기 | 계획 및 설계 | | 데이터 수집 및 처리 | | | | | | | 인공지능 모델 개발 | | | | | | | | 시스템 구현 | | | | | | | 운영 및 모니터링 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 요구사항 | 1 | | 2 | | 3 | | 4 | | | 5 | | 6 | 7 | 8 | | 9 | | 10 | 11 | | 12 | | | 13 | | | 14 | |
| | 세부항목 | 1-1 | 1-2 | 2-1 | 2-2 | 3-1 | 3-2 | 4-1 | 4-2 | 4-3 | 4-4 | 5-1 | 5-2 | 6-1 | 7-1 | 8-1 | 8-2 | 9-1 | 9-2 | 10-1 | 11-1 | 11-2 | 12-1 | 12-2 | 13-1 | 13-2 | 13-3 | 14-1 | 14-2 |
| Profile 대상여부 | | V | V | V | V | V | | | | | | V | V | | | V | V | | | | V | V | V | V | V | V | V | V | | |
| Profile 확인결과 | | Y | Y | Y | Y | Y | | | | | | Y | N | | | Y | Y | | | | Y | Y | Y | N | Y | Y | Y | Y | | |
| 권고사항 | | V | V | | V | V | | | | | | | | | | V | | | | | | | | | V | | | | | |

*검증항목 별 상세 결과는 '현장적용 결과' 엑셀시트 참조


## 핵심 요건별 준수율

**100% 다양성 존중** 관련 요구사항 4 6 10

인공지능이 특정 개인이나 그룹에 대한 차별적이고 편향된 관행을 학습하거나 결과를 출력하지 않으며, 모든 사람이 평등하게 B 솔루션의 혜택을 받을 수 있습니다.

**100% 책임성** 관련 요구사항 1 2 4 5 8 11 14

B 솔루션 개발 과정에서 위험에 대해 충분히 인지하고 있으며, 운영 과정에서 발생할 수 있는 사고에 대한 의사결정 추적 방안 및 운영 체계가 잘 수립되어 있습니다.

**80% 안전성** 관련 요구사항 3 5 7 11 13

데이터 수집 및 처리, 인공지능 모델 업데이트 및 개발, 운영 과정에서 발생할 수 있는 성능 저하 및 장애 대비를 위한 기술 구현, 내부 검수 및 운영 절차가 명확히 수립되어 있습니다.

**90% 투명성** 관련 요구사항 1 2 4 8 9 11 12 13 14

B솔루션의 기능 및 목적을 사용자에게 상세히 설명 가능한 매뉴얼이 마련되어 있으며, 고객사의 VoC 절차, 인터랙션을 고려한 화재 알림 평가 기준을 수립하여 신뢰할 수 있는 인공지능 서비스를 제공하고 있습니다.

## 권고사항

**1 계획 및 설계**
- 지속적이고 주기적인 인공지능 서비스 위험요소 도출 체계 및 원인 분석에 대한 자체 방법론 구축 관련 검증항목 : 01-1a
- 생명주기 전반에 걸친 위험 제거 활동 및 파급효과 감소 여부 확인을 위한 정량 혹은 정성적 평가방안 마련 관련 검증항목 : 01-2a

**2 데이터 수집 및 처리**
- 필요 시 End-user에게 오픈소스 데이터 출처 정보 제공에 대한 방법론 마련 관련 검증항목 : 04-2b
- 학습 데이터 이상값 검출을 위한 도구 또는 알고리즘의 필요성 논의 관련 검증항목 : 05-1b

**3 인공지능 모델 개발**
- 인공지능 모델 업데이트 시 내부 구성원의 모델 변경사항 공유를 위한 모델 명세(팩트 시트) 마련 관련 검증항목 : 10-2a

**4 시스템 구현**

해당 사항 없음

**5 운영 및 모니터링**
- 모델 변경에 따른 추적가능성 확보를 위한 추가, 변경 데이터 특징, 주요 시스템 변경 이력의 문서화 또는 이에 준하는 대처 방안 마련 관련 검증항목 : 14-2c