과제제안요구서(RFP)

과제명 : AI 적용 체계 안전설계기획 프레임워크 표준화 연구

1. 연구의 개요

가. 연구의 정의

본 연구는 자율주행차, 무인항공기, 무인수상정, 자율무기체계 등 AI 기반 자율시스템의 불확실성을 체계적으로 관리하기 위해 AI 안전설계 프레임워크와 AI 체계 안전성(안전경계결정) 시험평가 방법을 개발하고 표준화하는 것을 목적으로 한다.

나. 연구의 배경 및 필요성

o AI 자율체계의 불확실성 관리 필요

- AI는 학습 데이터 범위를 벗어난 환경에서 예측하기 어려운 판단을 수행할 수 있어 기존 안전성 검증체계만으로는 한계 존재

o 민·군 공통 안전설계 기준 부재

- 안전성 검토 및 검증이 결여된 상태로 현장에 도입될 경우, 예상치 못한 치명적 사고(민간인 피해 등)를 초래할 수 있으므로 객관적이고 구속력 있는 표준 가이드라인 필요
- 국방 분야에서는 AI의 오작동이 단순한 서비스 장애를 넘어 작전 실패, 전력 손실, 인명 피해 등으로 이어질 수 있으므로, 개발 완료 후 안전성을 검증하는 기존 방식만으로는 안전성 확보 한계 발생

o 책임 있는 AI(RAI, Responsible AI) 구현 요구 확대

- 최근 국제적으로 AI 시스템은 개발 이후 안전성을 평가하는 접근에서 벗어나, 기획·설계 단계부터 안전성을 내재화하는 “AI Safety by Design(안전설계)” 체계구축 중심의 연구 추진
- 미국 국방부의 자율무기체계 지침(DoDD 3000.09) 등 글로벌 안보 기준 등 국내외 정책은 인간의 통제권 보장과 위험관리 체계 확보를 요구하고 있으며 이를 기술적으로 구현할 표준 필요

o AI 경계결정 안전성 시험평가 표준화

- AI 자율시스템이 위험 상황에서 안전 경계를 설정하는 로직을 객관적으로 검증할 수 있는 표준 시험 절차와 평가지표 개발 필요
- 고위험·스트레스 시나리오 기반의 검증 기술 고도화(가혹 환경(Corner Case) 및 Edge Case 도출) 필요

다. 연구 최종 목표

o AI 적용 체계 안전설계 프레임워크 표준(안) 개발

* ① 위험 식별(Risk Identification), ② 위험 분석(Risk Assessment), ③ 안전 요구사항 정의(Safety Requirements), ④ 안전설계 및 개발(Safety Engineering), ⑤ 검증 및 시험평가(Verification & Validation), ⑥ 운영 및 모니터링(Operation Monitoring), ⑦ 지속 개선(Continuous Improvement) 등 무기체계 전순기 프레임워크 중심 표준(안) 마련

o AI 위험관리 체계 수립

o 인간개입(HITL) 및 최소위험상태 기준 정립

o AI 안전성(경계결정) 시험평가 절차 표준(안) 개발

o 국제표준 연계 국방표준화 기반 마련

2. 연구 현황 및 전망

가. 국내

o 국내 AI 안전성 연구는 자율주행차, 무인기, 국방 AI, 산업용 로봇 분야를 중심으로 확대되고 있으며, 최근에는 단순 성능 검증을 넘어 AI의 신뢰성·안전성 확보를 위한 시험평가 체계 구축 연구가 추진되고 있음

o 국토교통부, 자동차안전연구원(KATRI), 한국전자기술연구원(KETI), 한국전자통신연구원(ETRI) 등을 중심으로 자율주행차 시험평가 및 안전성평가 연구 위주로 진행

o 운용요구서(ORD), 운용형태·임무유형(OMS/MP)과 연계된 ODD 정의, 인간개입(HITL) 설계, AI 안전 경계결정기능을 통합적으로 다루는 표준화 연구는 부재

o AI 윤리 및 책임성(Responsible AI) 논의가 확대되고 있으나, 실제 획득·인증·전력화 과정에서 활용 가능한 공학적 안전설계 및 시험평가 절차 표준으로 연결은 제한되는 실정임

나. 국외

o 미국은 국방부(DoD)와 CDAO를 중심으로 자율무기체계 안전성 확보를 위한 정책과 기술 연구를 선도하고 있으며, DoDD 3000.09를 통해 인간의 적절한 판단과 통제(Appropriate Levels of Human Judgment)를 핵심 원칙으로 규정

o 최근 국제적으로 Responsible AI Strategy를 수립하여 신뢰성(Reliability), 추적가

능성(Traceability), 통제가능성(Governability) 등을 핵심 요구사항으로 제시하고 있으며, 회원국 간 AI 안전성 평가체계 공동 연구를 확대

o 국제적으로도 소요기획 단계부터 ODD 정의, 위험 분석, HITL 설계, AI 경계결정, 가상·실물 통합 검증 및 시험절차 표준화 등 전 수명주기 프레임워크 표준화사례는 제한적임

다. 국내외 연구수준 비교 및 협력 가능성

o 국내는 자율주행, 디지털 트윈, AI 신뢰성 평가 등 개별 기술 분야에서 세계 선진국 대비 약 70~80% 수준의 기술력을 확보한 것으로 평가되나, AI 안전설계와 시험평가를 통합하는 체계공학(System Engineering) 기반 프레임워크 연구는 아직 초기 단계임

o 미국·EU는 AI 위험관리, 인간 통제(HITL), 책임 있는 AI(RAI) 정책과 제도 분야에서 선도적 위치를 확보하고 있으며, 국제표준화 활동에서도 주도권을 보유하고 있으며, 향후 ISO/IEC, SAE 등 국제 표준화 기구와의 협력을 통해 국내 연구성과와 국제표준과의 연계할 필요성이 있음

o 국내외에서 개별적으로 수행되고 있는 ODD, Responsible AI, AI Risk Management 연구 등을 분석하고, 민·군 공통의 AI 안전설계 및 시험평가 표준 프레임워크 구축 필요

3. 연구개발계획

가. 연구 목표

o 국방 AI 안전설계 개념 및 범위 정립
- 국내외 AI 안전 관련 정책, 표준, 프레임워크 분석
- 국방 환경에 적합한 AI Safety by Design 개념 정의
- 국방 AI 특화 위험요소 및 안전속성 도출

o AI 위험기반 안전설계 프로세스 개발
- AI 생애주기 기반 위험관리 절차 수립
- 위험 식별(Risk Identification) 및 위험평가(Risk Assessment) 체계 개발
- 안전 요구사항 도출 방법론 정립
- 최소위험상태(Minimal Risk Condition) 및 인간개입(HITL) 적용 기준 마련

o 국방 AI 안전설계기획 프레임워크 개발
- 안전설계 활동, 역할, 책임, 산출물 정의
- 기획단계 안전성 검토체계 개발
- 시험평가 및 운영단계 연계체계 구축
- AI 신뢰성·안전성 확보를 위한 표준 프로세스 제시
- Agile 환경과 연계 가능한 안전설계 체계 수립

o 안전경계결정 및 시험평가 표준(안) 정립

- AI에 치명적인 스트레스를 유발하는 고위험 시나리오를 도출하고, 설계-검증을 연결하는 양방향 파라미터 최적화
- 인간 개입 설계 및 AI 안전 경계를 식별
- 위협 조건 발생 시 인간과 AI 간의 제어권 전환 및 시스템의 최소위험상태 진입 규칙 설계
- AI 안전성(안전경계결정) 시험평가 표준(안) 마련

나. 연구 내용

구분	연구내용
1차년도	· 국내외 AI 안전설계 및 안전성 검증 표준화 동향 조사 · ORD·OMS/MP 기반 AI 운용설계영역(ODD) 정의 · 지상·공중·해상 도메인별 ODD 분류체계 수립 · 국방 AI 위험요소 식별 및 AI 안전설계 범위 및 용어체계 정립 · FMEA/STPA 기반 하이브리드 위험분석 체계 분석 · AI 윤리 및 책임성(Responsible AI) 확보를 위한 국제규제 분석 및 정합성 확보
2차년도	· AI 안전성 경계(Safety Boundary) 개념 연구 · 최소위험상태(MRC) 전환 로직 설계 · Edge/Corner Case 기반 스트레스 시나리오 구축 · 생성형 AI 활용 위험 시나리오 자동 생성 · AI 안전성 평가 메트릭 개발 · 안전경계결정 개념 및 모델링
3차년도	· 안전설계기획 프레임워크 표준(초안) 및 안전성 시험평가 표준(초안) 개발 · 인간개입(HITL) 및 통제권 전환 규칙 개발 · SW-in-the-Loop(SIL) 검증 수행 · ODD 기반 가상 스트레스 테스트 수행 · 맥락 충분성 검토 및 알고리즘 보정
4차년도	· UGV·UAS 실물 대상 시험평가 표준(안) 검증 · 인간개입(HITL) 및 안전경계 기능 실증 · Sim-to-Real Gap 분석 및 성능 보정 · 민·군 적용성 검증 및 실증 데이터 확보
5차년도	· AI 안전설계 프레임워크 표준(안) 정립 · 민·군 공통 안전성 시험평가 표준(안) 정립 · ODD 기반 안전성 평가 가이드라인 작성 · 국방표준서 제정(안) 의견수렴 및 전문가 검토 회의

다. 전체 연구개발비 최대 지원규모 : 10억원 (연구개발기간 : 48개월)

- 소요예산은 R&D 예산 편성에 따라 변경될 수 있음

4. 적용 및 파급효과

가. 적용분야

o 민수

- 자율주행차, UAM, MASS 등 고위험 모빌리티 분야의 AI 안전성 검증 및 인증 체계 구축에 활용
- 스마트팩토리·산업용 로봇 분야에서 인간-로봇 협업(HRI) 환경의 안전성 확보 및 위험상황 대응기준 마련
- 철도, 원자력, 에너지, 항만 등 국가 핵심 인프라의 AI 기반 자동화 시스템 안전성 검증에 적용
- 자율시스템 개발기업이 설계단계부터 안전성을 반영할 수 있는 공통 프레임워크 제공

o 군수

- 무인기(UAV), 무인수상정(USV), 무인차량(UGV) 등 AI 기반 무기체계의 안전성 확보 및 전력화 지원
- 유무인복합체계(MUM-T) 운용 시 인간개입(HITL) 및 통제권 전환 기준 제공
- GPS 교란, 전파방해, 악천후 등 전장환경에서 AI 시스템의 신뢰성 검증체계 구축
- ORD·OMS/MP 기반 운용설계영역(ODD) 정의를 통해 획득 초기단계부터 위험요소를 체계적으로 관리
- AI 시스템 객관적 평가기준 마련 및 획득 리스크 감소

나. 파급효과

o 기술적 측면

- ODD, HITL을 연계한 AI 안전설계 방법론 확립으로 국내 AI 안전 기술 수준 향상
- 설계 - 검증 - 운용을 연계하는 전주기 안전성 관리체계 구축
- 디지털 트윈, SIL/HILS, 스트레스 테스트 기반의 고도화된 AI 시험평가 기술 확보
- AI 불확실성, 강건성, 신뢰성에 대한 정량평가 기술 확보 및 객관적 검증체계 정립
- 민·군 공통 활용이 가능한 AI 안전설계 및 시험평가 표준 기반 마련으로 국가 AI 신뢰성 경쟁력 강화

o 경제적 측면

- AI 자율시스템 개발과정에서 반복적인 재설계 및 재시험 비용을 절감하여 개발 기간과 사업비를 절감
- 민수·군수 공통 활용이 가능한 표준 프레임워크 구축으로 중복투자를 방지하고 연구개발 효율성 향상
- 국내 AI 안전성 시험·인증 산업 육성을 통한 신규 시장 창출 및 관련 전문인력 수요 확대

- 자율주행, 로봇, 드론 산업의 사업화 촉진과 글로벌 시장 진출 기반 확보
- 사고 예방을 통한 사회적 비용 감소 및 AI 기반 산업 전반의 투자 활성화 기대

5. 연구 결과 제시물 및 평가항목

가. 연구결과 보고서 및 표준(안)

o 최종보고서
- 소요기획 연계형 AI 적용 체계 운용설계영역(ODD) 정의서 및 표준 양식
- 위험 기반 안전조치 규칙 설계 지침
- 인간 개입(HITL) 및 맥락 충분성 검토를 포함한 안전성 평가 절차서
- 표준 적용 가이드라인 (소요기획 및 민수/군수 활용 실무 가이드)

o 표준(안)
- AI 적용 체계 안전설계 프레임워크 통합 표준
- AI 안전성(경계결정) 시험평가 표준

나. 평가항목

o 연구 수행방법 및 과정의 타당성
o 최종 목표의 달성도
o 연구결과의 활용성(민·군 분야에서의 이용 가능성) 등

6. 참여 요건

가. 추진 체계 요건

o 주관연구개발기관 : 민·군기술협력사업 촉진법 제7조제2항 및 동법 시행령 제14조제2항 각 호에 해당하는 기관 또는 단체
o 공동 및 위탁연구개발기관 : 제한 없음 (기업참여의 경우 참여 필요성 및 역할 제시)
o 기업 분담율 : 국가연구개발혁신법 시행령 제19조

나. 연구책임자의 자격 및 과제 신청요건

o 연구책임자의 자격 : 관련분야의 연구 경험이 풍부한 연구자를 책임자로 선임하여 연구의 최종목표를 달성할 수 있도록 계획, 업무프로세스 정립, 원활한 추진 및 조정과 과제관리를 수행할 수 있어야 함
o 과제 신청요건 : 주관연구개발기관은 제안한 연구개발 목표를 충분히 달성할 수 있는 연구팀을 구성하여야 하며, 필요시 컨소시엄을 구성할 수 있음