1. AI Safety Framework

소비자 가전 제품(CE)이 온디바이스 AI 기능을 점점 더 통합해 감에 따라, AI Safety 는 매우 중요한 문제로 대두되었습니다. AI Safety 와 관련된 문제를 해결하기 위해 삼성전자는 2 가지 핵심 프로세스와 4 가지 핵심 컴포넌트에 초점을 맞춘 포괄적인 온디바이스 AI Safety Framework 를 구축했습니다.

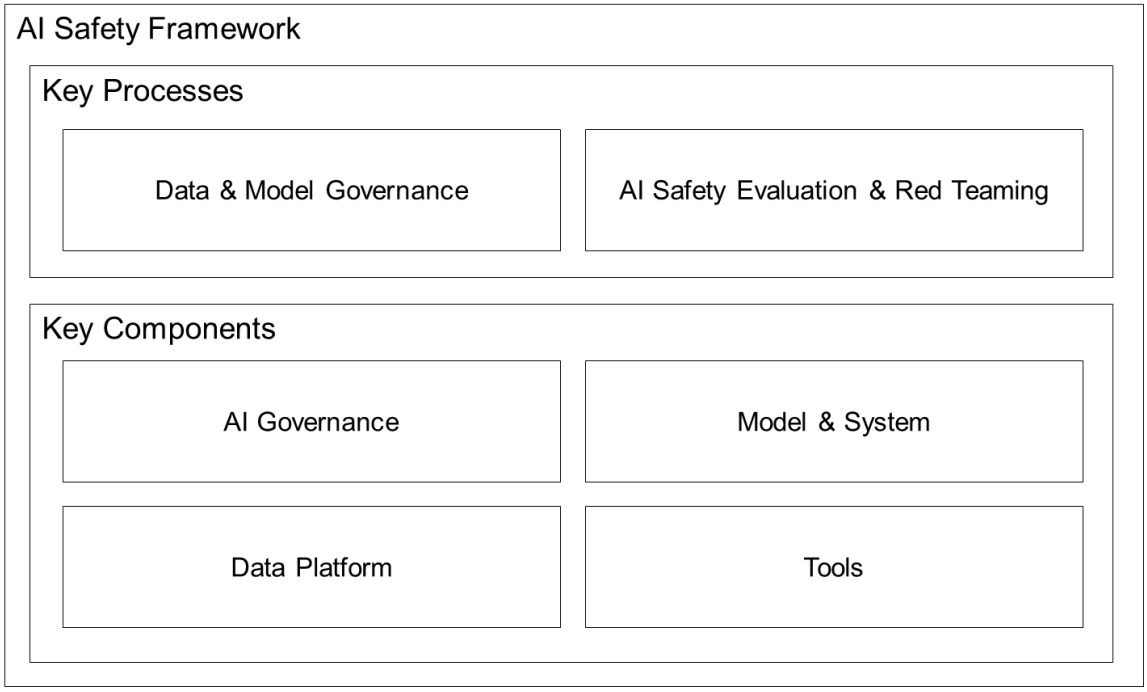

Figure 1 AI Safety Framework 구조

제품에서 AI Safety Framework 의 중요성

온디바이스 AI 의 사용자 프라이버시 보호, AI 신뢰성 보장, 적대적 공격에 대한 잠재 위험 완화 등을 위해서, AI Safety 확보는 중요합니다. 삼성전자의 AI Safety Framework 은 공정성, 투명성 및 책임성을 유지함으로써 이러한 위협을 총체적으로 관리하는 것을 목표로 합니다. 사용자 신뢰를 유지하고 국제 표준을 준수하기 위해 구조화된 Framework 를 통해 안전성을 보장하고자 합니다.

AI Safety 확보를 위한 주요 프로세스

AI Safety Framework 은 다음과 같은 프로세스를 포함합니다.

Data & Model Governance

데이터를 획득하거나 수집할 때, 데이터의 품질과 적절성을 평가하고 잠재적인 프라이버시 및 기타의 예상되는 문제를 해결하기 위한 작업을 수행합니다. 데이터 획득 프로세스는 (i) 데이터의 특성을 식별하고 위험 및 품질 평가를 수행하는 것을 포함하는 데이터 사양 계획 및 검증 단계, (ii) 데이터 획득 절차를 면밀히 모니터링하는 단계, (iii) 데이터 정보를 기록하는 것을 포함하는 데이터 검증 단계, (iv) 데이터 등록의 4 개 컴포넌트로 구성됩니다. 데이터 카드를 활용하여 등록된 데이터를 관리하고 추적합니다.

또한 AI 모델에 대한 위협을 파악하고 완화하기 위해서 모델 카드 작성 및 모델 평가를 진행하고, Safety 원칙 체크리스트를 통해 공정성, 투명성, 책임성 평가를 진행합니다.

AI Safety Evaluation & Red Teaming

각 AI 모델에 대해, 잠재적 위험을 파악하고 완화하기 위한 Safety 평가를 수행합니다. 이 평가는 공개된 Safety 벤치마크를 활용하여 유해성, 편향성, 진실성, 전반적인 신뢰성 등 다양한 측면에서 Safety 위험을 정량적으로 평가하여 의도하지 않은 결과를 방지합니다.

또한 통제된 적대적 기법을 사용해 잠재적인 보안 위협을 시뮬레이션하여 모델에 대한 공격을 방어하는 능력을 강화하고 실제 위험에 더 잘 대비할 수 있도록 합니다.

이러한 위험을 선제적으로 해결하기 위해, AI 모델 관련 Safety 위협을 식별하고 해결하는데 중점을 둔 AI Red Teaming 프로세스를 개발하였습니다.

AI Safety 평가 및 Red Teaming 프로세스는 기존에 회사에서 활용하고 있는 보안 개발 라이프사이클을 기반으로 4 단계로 구성되어 운영됩니다.

Planning: AI 모델 및 시스템 개발자와의 내부 검토를 통해 잠재적인 AI 위협 모델을 결정하고 그에 따른 평가를 계획합니다. 명확한 목표 설정, 범위 정의, 잠재적 위험 파악, 효과적인 테스트에 필요한 자원 및 전략 수립에 초점을 맞춥니다.

Analysis & Design: 대상 AI 모델에 대한 위험도 평가를 실시하고, 발생 가능한 위험도를 파악하여 Safety 평가의 요구사항을 추출합니다. 위험도 평가를 기반으로 적절한 AI Safety Benchmark 데이터셋과 참조 AI 모델에 대한 결정을 포함하는 경험적 테스트를 설계합니다. 또한, Red Teaming 을 위해 보안 요구사항을 철저히 분석하고, 테스트 중인 시스템을 점검하며, 위협 모델링을 수행하고, 적대적인 테스트를 설계하여 잠재적 취약점을 파악합니다.

Evaluation & Testing: 대상 AI 모델에 대한 평가를 실시하고 참조 모델과 비교합니다. 또한 실제 제품 개발/생산 시스템에 대한 잠재적인 간섭을 방지하기 위해 분리되고 격리된 Red Team 테스트 환경을 설정했고, 수동 및 자동화 테스트를 모두 실행합니다. 중간 보고서에 대한 AI 개발팀의 피드백을 수집하여 테스트 목표와 결과가 잘 맞지 않을 때 이에 대한 조치를 수행합니다.

Operation & Maintenance: 최종 검증 보고서가 생성된 후, 평가팀은 확인된 취약점이 신속하고 효과적으로 해결되도록 재평가 계획을 수립합니다. 개발팀은 재평가를 위해 모델을 재학습하거나 보다 안전한 다른 모델을 선택하는 등 가능한 완화 방안을 검토하고 가장 적합한 모델을 선정합니다.

이러한 포괄적인 구조는 AI 기술 및 서비스의 제공에 있어서 사용자 및 회사를 보호할 수 있는 신뢰 환경을 마련하기 위한 접근입니다.

AI Safety 프로세스를 지원하기 위한 주요 컴포넌트

AI Safety 프로세스를 지원하기 위한 4 개의 주요 컴포넌트가 포함됩니다.

Model & System: AI 모델이 Safety 기준에 맞게 설계되고 관리될 수 있도록 관리합니다. 이 컴포넌트에는 AI 모델 개발 및 라이프사이클 전반에 걸친 주기적 모델 평가 및 검증 프로세스 수행이 포함됩니다.

Data Platform: 학습 데이터의 관리 방안과 AI 모델 배치 등 안전한 데이터 처리를 지원합니다.

Tools: AI 모델을 안전하게 운영 관리, 모니터링, 평가하고 지원하기 위해 통합된 기술 자원을 제공합니다.

AI Governance: 데이터 및 모델의 사용, 제어 및 관리에 대한 내부 정책과 거버넌스를 유지/관리합니다.

거버넌스 운영

삼성전자는 AI Safety Framework 실행 모니터링 및 관리를 위해 운영 조직인 AI 전략팀을 신설하였습니다.

AI 전략팀은 다양한 관련 부서와 연계하여 데이터, 모델 수명 주기, AI Safety 관련 프로세스를 관리하고 관련 법과 규제에 따라 프로세스를 업데이트합니다.