정책이슈리포트 2025-07

영국 AI 사이버 보안 실행 강령

AI Cyber Security Code of Practice 주요 내용

발행일: 2025.07.04.
발행기관: 디지털안전센터
키워드: 사이버보안, AI안전, 행동규범

1. 배경

영국이 2025년 1월 31일 AI 시스템의 사이버보안 위험 해결을 위한 자발적 실행규범「AI 사이버 보안 실행 강령」을 발표했다.

영국 과학혁신기술부(DSIT)와 국가사이버보안센터(NCSC)가 2025년 1월 31일 발표한「AI 사이버 보안 실행 강령(AI Cyber Security Code of Practice)」은 AI 시스템의 사이버보안 위험 해결을 위한 혁신적인 자발적 실행규범이다. 발표와 동시에 즉시 적용 가능한 이 강령은 'Secure by Design' 원칙을 AI 영역에 적용하여 설계 단계부터 폐기 단계까지 AI 시스템의 전체 생명주기에 걸친 13개의 핵심 보안 원칙을 제시한다. 데이터 오염, 모델 역전, 멤버십 추론 등 기존 소프트웨어와 차별화되는 AI 특화 위협에 대한 구체적 대응 방안을 담고 있다. 영국 정부는 이 실행강령을 기반으로 유럽전기통신표준연구소(ETSI)를 통해 글로벌 표준(TS 104 223)을 개발할 예정이며, 이는 전 세계 AI 보안 표준화를 선도하는 중요한 이정표가 될 것으로 평가된다.

2. 의견수렴 과정 및 결과

2.1. 의견수렴 과정

영국 정부는 2024년 5월 15일부터 8월 9일까지 DSIT는 AI 사이버보안에 대한 의견수렴을 실시했다. 이 과정에서 총 123건의 응답을 받았으며, 응답자의 68%가 개인이 아닌 기관이었다. 글로벌 의견수렴으로 진행되어 영국, 유럽, 미국, 일본, 싱가포르 등 다양한 국가에서 응답을 받았다.

응답 현황

- 총 응답 수: 123건
- 기관 응답 비율: 68%
- 지역별 분포: 영국(54%), 북미(26%), 아시아(11%), 유럽(7%)

2.2. 주요 피드백 및 개선사항

응답자의 80%가 이러한 실행규범 마련을 지지했으며, 실행규범의 각 원칙에 대해서는 83%에서 90% 범위의 높은 지지율을 보였다.

항목	지지율	반대율	모름
2단계 접근법 전체	80%	11%	9%
12개 개별 원칙	83-90%	5-9%	6-8%
이해관계자 구분	62%	29%	8%

의견수렴 과정에서 나타난 주요 피드백은 다음과 같다.

- 구현 가이드 필요성: 실행강령 구현을 위한 더 상세한 지침 요구
- 이해관계자 정의 개선: "데이터 관리자"를 "데이터 보관자"로 변경, "영향받는 개체" 추가
- 13번째 원칙 추가: AI 시스템의 생명주기 종료에 대한 원칙 신설
- 비용 및 역량 우려: 응답자의 62%가 재정적 영향이 있을 것으로 예상

의견수렴 결과 응답 기관의 62%가 실행강령 구현 시 재정적 영향이 있을 것으로 예상한다고 답했으나, 구체적인 비용 데이터를 제공한 응답자는 매우 적었으며, 대부분 자원과 광범위한 재정적 영향에 초점을 맞췄다. 영국정부는 보안 요구사항 구현에 따른 비용이 사이버 공격으로 인한 피해 비용보다 적다고 판단한다. 사이버 공격 성공 시 발생하는 금전적 손실, 데이터 손실, 평판 손상 등을 고려할 때 사전 보안 투자가 더 경제적이라는 입장이다.

정부는 AI 사이버보안 분야의 기술 격차에 대한 우려가 제기됨에 따라, AI사이버보안 구현을 위한 상세한 가이드(실행 지침) 마련, 사이버보안 위원회 지원, 시장조사 실시, 사이버보안 인력 양성 지원(CyberFirst 프로그램) 등을 마련했다.

이번 의견수렴 결과로 이해관계자에 대한 정의와 데이터 관리자에 대한 개념이 보다 명확해졌다. 이해관계자의 범위를 영향받는 개체(Affected Entities)로 확대함으로써 이해관계자의 범위를 확대했다.

이해관계자	역할	변경사항
개발자 (Developer)	AI 시스템 설계, 구축, 훈련, 적응	오픈소스 모델 적응자 포함
시스템 운영자 (System Operator)	AI 시스템 운영 및 관리	기존 정의 유지
데이터 보관자 (Data Custodian)	데이터 정책 수립 및 관리	"데이터 관리자"에서 변경
최종 사용자 (End-user)	AI 시스템 직접 사용	NIST 정의와 일치하도록 수정
영향받는 개체 (Affected Entities)	AI 시스템 결정에 간접 영향	신규 추가

또한, 의견수렴 결과를 반영하여, AI 시스템의 생명주기 종료에 대한 13번째 원칙이 새롭게 추가되었으며, 제품의 생명주기 종료에 따라 훈련 데이터 및 모델의 소유권 이전과 시스템 폐기에 관한 안전한 조치를 다룬다.

정부는 이러한 피드백을 반영하여 실행강령을 업데이트하고 새로운 구현 가이드를 작성했으며, 이를 ETSI 글로벌 표준 개발의 기반으로 활용할 예정이다.

3. 목적 및 주요 내용

3.1. 목적

이 규범의 주요 목적은 AI 시스템의 사이버보안 위험 해결을 위한 자발적 실행규범을 제시하여, 유럽전기통신표준연구소(ETSI)의 글로벌 표준 수립을 지원하는 것이다. 또한 AI 공급망 전반의 이해관계자들에게 AI 시스템 보호를 위해 구현해야 할 기준 보안 요건에 대한 명확한 지침을 제공하는 것을 목적으로 한다.

3.2. 자발적 권고 성격

자발적 성격을 띠며 강제성보다는 권고의 형태로 구성된다. 각 조항은 "Shall"(요건), "Should"(권고사항), "May"(가능성)로 구분하여 의무 수준을 명확히 하며, 기존 소프트웨어 실행규범의 부록으로 위치하여 모듈화된 접근을 취하고, 실무 적용을 위한 상세한 구현 가이드를 함께 제공한다.

구분	의미	의무 수준
Shall	요건	필수 준수
Should	권고사항	권장 준수
May	가능성	선택적 준수

3.3. 적용범위와 대상

이 규범은 심층 신경망을 포함하는 AI 시스템 전반에 적용되며, 특히 생성형 AI와 같은 고도화된 AI 기술들을 주요 대상으로 한다. 다만, 연구 목적으로만 생성되고 테스트되어 실제 배포되지 않을 AI시스템들은 규범의 적용대상에서 제외된다.

이해관계자	적용범위와 주요 책임
개발자	• 범위: 독점 모델과 오픈소스 모델을 모두 포함 • 책임: 모델 설계, 개발, 테스트, 문서화
시스템 운영자	• 범위: 직접운영 및 서비스 제공업체를 모두 포함 • 책임: 배포, 운영, 모니터링, 최종사용자 지원
데이터 관리자	• 범위: 데이터 정책설정 및 관리권한 보유자 • 책임: 데이터 보안, 접근 통제, 정책 수립
최종사용자	• 범위: 업무 및 일상활동 지원 목적의 사용자 • 책임: 적절한 사용, 보안 위험 신고
영향을 받는 개체	• 범위: 직접 AI와 상호작용이 없어도 영향을 받을 수 있는 잠재적 대상자 모두 • 책임: 권리 보호, 투명성 요구

3.4. 보안활동 요구사항

실행규범은 데이터 오염, 모델 역전, 멤버십 추론과 같은 AI 특화 위협에 있어 AI 시스템의 전체 생명주기를 아우르는 포괄적 접근을 통해 설계 단계부터 폐기 단계까지 각 단계별로 요구되는 보안활동 총 13개 원칙을 제시한다.

생명주기 단계	13개 원칙
설계단계	#1 AI 보안 위협과 위험에 대한 인식 제고
	#2 기능성과 성능뿐만 아니라 보안을 위한 AI 시스템 설계
	#3 위협 평가 및 AI 시스템 위험 관리
	#4 AI 시스템에 대한 인간의 책임 활성화
개발단계	#5 자산 식별, 추적 및 보호

	#6 인프라 보안
	#7 공급망 보안
	#8 데이터, 모델 및 프롬프트 문서화
	#9 적절한 테스트 및 평가 수행
배포 및 유지보수	#10 최종사용자 및 영향을 받는 개체와 관련된 커뮤니케이션 및 프로세스
	#11 정기적인 보안 업데이트, 패치 및 완화 조치 유지
	#12 시스템 행동 모니터링
	#13 적절한 데이터 및 모델 폐기 보장

1) 설계단계(원칙 1~4): AI 시스템을 기획하고 설계할 때 보안을 처음부터 고려하는 단계

원칙	주요 책임자	핵심 보안활동
원칙 1: AI 보안 위협 및 위험 인식 제고	시스템 운영자, 개발자, 데이터 관리자	• 정기적인 AI 보안 교육 프로그램 운영 • 직무별 맞춤형 보안 교육 제공 • 최신 AI 보안 위협 동향 공유 • 개발자 대상 AI 보안 코딩 교육 실시
원칙 2: 보안을 고려한 AI 시스템 설계	시스템 운영자, 개발자	• 비즈니스 요구사항 및 보안 위험 평가 • 적대적 AI 공격 대응 설계 • 외부 컴포넌트 보안 위험 평가 • 시스템 간 연동 시 최소 권한 원칙 적용 • 외부 공급업체 실사 수행
원칙 3: 위협 평가 및 위험 관리	시스템 운영자, 개발자	• 정기적인 위협 모델링 수행 • AI 특화 공격(데이터 포이즈닝, 모델 역추적 등) 대응 • 위험 기반 보안 통제 적용 • 외부 위협 정보 공유 및 전달 • 지속적인 시스템 모니터링
원칙 4: AI 시스템에 대한 인간 책임 구현	시스템 운영자, 개발자	• 인간 감독 기능 설계 및 유지 • 모델 출력 해석 가능성 확보 • 인간 감독 기반 위험 통제 설계 • 금지 사용 사례 명시 및 안내

2) 개발단계(원칙 5~9): AI 시스템을 실제로 구축하고 개발하는 과정에서 보안을 구현하는 단계

원칙	주요 책임자	핵심 보안활동
원칙 5: 자산 식별, 추적 및 보호	시스템 운영자, 개발자, 데이터 관리자	• 포괄적인 자산 인벤토리 관리 • 버전 관리 및 자산 보안 도구 운영 • AI 시스템 특화 재해 복구 계획 수립 • 민감 데이터 무단 접근 차단 • 데이터 및 입력값 검증 및 정제

원칙 6: 인프라 보안 강화	시스템 운영자, 개발자	• API, 모델, 데이터 접근 통제 • 외부 API 공격 완화 통제 • 개발/운영 환경 분리 • 취약점 공개 정책 수립 • 사고 관리 및 복구 계획 운영
원칙 7: 공급망 보안 확보	시스템 운영자, 개발자, 데이터 관리자	• 보안 소프트웨어 공급망 프로세스 적용 • 외부 모델/컴포넌트 사용 정당성 문서화 • 완화 통제 및 위험 평가 수행 • 모델 업데이트 시 사전 공지
원칙 8: 데이터, 모델, 프롬프트 문서화	개발자	• 시스템 설계 및 유지보수 계획 문서화 • 보안 관련 정보 포함 문서 작성 • 모델 컴포넌트 암호화 해시 제공 • 공개 데이터 출처 및 수집 정보 기록 • 시스템 프롬프트 변경 감사 로그 유지
원칙 9: 적절한 테스트 및 평가 수행	시스템 운영자, 개발자	• 보안 평가 프로세스 내 모델 테스트 • 배포 전 독립적인 보안 테스트 • 테스트 결과 공유 및 협업 • 모델 출력 역공학 방지 평가 • 의도하지 않은 시스템 영향 평가

3) 배포 및 유지보수 단계(원칙 10~13): AI 시스템을 실제 환경에 배포하고 지속적으로 안전하게 운영하는 단계

원칙	주요 책임자	핵심 보안활동
원칙 10: 최종 사용자 및 영향받는 개체와의 소통	시스템 운영자	• 데이터 사용/접근/저장 방식 투명하게 공개 • 시스템 사용/관리/설정 가이드 제공 • 모델 제한사항 및 잠재적 실패 모드 안내 • 보안 관련 업데이트 사전 공지 • 사이버 보안 사고 시 사용자 지원
원칙 11: 정기적인 보안 업데이트 및 패치	개발자, 시스템 운영자	• 보안 업데이트 및 패치 제공 • 업데이트 불가능한 경우 완화 방안 마련 • 주요 업데이트 시 새로운 보안 테스트 수행 • 모델 변경 평가 및 대응 지원
원칙 12: 시스템 동작 모니터링	개발자, 시스템 운영자	• 시스템 및 사용자 행동 로그 기록 • 로그 분석을 통한 이상 행동 탐지 • AI 시스템 내부 상태 모니터링 • 모델 성능 변화 추적 및 분석
원칙 13: 적절한 데이터 및 모델 폐기	개발자, 시스템 운영자	• 훈련 데이터/모델 소유권 이전 시 보안 폐기 • 모델/시스템 폐기 시 데이터 및 설정 정보 안전 삭제 • 데이터 관리자 참여 하에 폐기 절차 수행

4. 향후 일정

정부는 향후 ETSI 글로벌 표준 및 가이드를 반영하여 규정 및 가이드의 내용을 업데이트할 예정이다. DSIT는 NCSC와 협력하여 규정 및 구현 가이드를 유럽 전기통신표준협회(ETSI)에 제출할 예정이며, ETSI는 이를 새로운 글로벌 표준(TS 104 223) 및 관련 구현 가이드(TR 104 128)의 기반으로 활용할 계획이다.

영국은 ETSI 외에도 ISO, CEN-CENELEC, ITU 등 다양한 표준화 기구와의 협력을 지속하고 있다. 또한 미국과 유럽 파트너들과의 협력을 통해 국제적으로 일치된 보안 요구사항 개발을 추진하고 있다.

5. 정책적 시사점

영국의 AI 사이버 보안 실행 강령은 AI 기술의 안전한 발전을 위한 선제적 접근을 보여주며, 'Secure by Design' 원칙을 AI 영역에 적용한 혁신적 시도로 평가된다.

영국의 「AI 사이버 보안 실행 강령」에 기반한 국제 협력을 통한 글로벌 표준화 추진은 ETSI 글로벌 표준(TS 104 223) 및 구현 가이드(TR 104 128)의 기반이 되어 전 세계 AI 보안 표준화를 주도할 것으로 예상된다. 이는 AI 기술 발전과 보안 강화를 동시에 추구하는 균형잡힌 접근법으로서, 향후 전 세계 AI 보안 정책 수립에 중요한 참고 사례가 될 것으로 평가된다.

참고문헌

Department for Science, Innovation & Technology (2025. 1. 31.). GuidanceCode of Practice for the Cyber Security of AI, https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice/code-of-practice-for-the-cyber-security-of-ai

Department for Science, Innovation & Technology (2025. 1. 31.). Call for evidence outcome Government response on AI cyber security, https://www.gov.uk/government/calls-for-evidence/cyber-security-of-ai-a-call-for-views/outcome/government-response-on-the-cyber-security-of-ai

별첨. 영국 AI사이버보안 실행강령

* 하단의 번역은 참고용으로 정확한 법적 해석은 원문 확인 필요.
원문출처: https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice/code-of-practice-for-the-cyber-security-of-ai

AI 사이버보안 실행강령(원문)	AI 사이버보안 실행강령(번역)
Introduction The UK government is taking forward a two-part intervention to address the cyber security risks to AI. This involves the development of a voluntary Code of Practice which will be used to help create a global standard in the European Telecommunication Standards Institute (ETSI) that sets baseline security requirements. We believe a Code focused specifically on the cyber security of AI is needed because AI has distinct differences to software. These include security risks from data poisoning, model obfuscation, indirect prompt injection and operational differences associated with data management. Further examples of the unique risks posed by AI systems can be found in Appendix B within the National Institute of Standards and Technology's (NIST) Risk Management Framework. The government is also intervening in this area because software needs to be secure by design and stakeholders in the AI supply chain require clarity on what baseline security requirements they should implement to protect AI systems. The proposed intervention was endorsed by 80% of respondents to the Department for Science, Innovation and Technology's (DSIT) Call for Views which was held from 15 May to 9 August 2024. Support for each principle in the Code ranged from 83% to 90%. This document also builds on NCSC's Guidelines for Secure AI Development which were published in November 2023 and endorsed by 19 international partners. As set out in DSIT's modular approach to cyber security codes of practice, AI stakeholders should view this document as an addendum to the Software Code of Practice.	서론 영국 정부는 AI의 사이버보안 위험을 해결하기 위해 두 부분으로 구성된 개입을 추진하고 있다. 이는 기준 보안 요건을 설정하는 유럽전기통신표준연구소(ETSI)의 글로벌 표준 수립에 활용될 자발적 실행규범의 개발을 포함한다. AI에 특화된 사이버보안 실행규범이 필요한 이유는 AI가 소프트웨어와 뚜렷한 차이점이 있기 때문이다. 이러한 차이점에는 데이터 오염, 모델 난독화, 간접적 프롬프트 주입으로 인한 보안 위험과 데이터 관리와 관련된 운영상의 차이점이 포함된다. AI 시스템이 제기하는 고유한 위험의 추가적인 예시는 미국 국립표준기술연구소(NIST)의 위험관리 프레임워크 부록 B에서 찾을 수 있다. 정부가 이 분야에 개입하는 이유는 소프트웨어가 설계부터 안전해야 하고, AI 공급망의 이해관계자들이 AI 시스템을 보호하기 위해 구현해야 할 기준 보안 요건에 대한 명확성을 필요로 하기 때문이다. 제안된 개입은 2024년 5월 15일부터 8월 9일까지 실시된 과학혁신기술부(DSIT)의 의견수렴에서 응답자의 80%가 지지했다. 실행규범의 각 원칙에 대한 지지율은 83%에서 90% 범위였다. 이 문서는 또한 2023년 11월에 발표되어 19개 국제 파트너의 지지를 받은 NCSC의 [안전한 AI 개발을 위한 가이드라인]을 기반으로 한다. DSIT의 [사이버보안 실행규범 모듈화 접근법]에 명시된 바와 같이, AI 이해관계자들은 이 문서를 소프트웨어 실행규범의 부록으로 간주해야 한다.
Scope The scope of this voluntary Code of Practice is focused on AI systems. This includes systems that incorporate deep neural networks, such as generative AI. For consistency, we have used the term "AI systems" throughout the document when framing the scope of provisions and "AI security" which is considered a subset of cyber security. The Code is not designed for academics who are creating and testing AI systems only for research purposes (AI systems which are not	적용 범위 이 자발적 실행규범의 적용 범위는 AI 시스템에 초점을 맞추고 있다. 여기에는 생성형 AI와 같은 심층 신경망을 포함하는 시스템이 포함된다. 일관성을 위해, 조항의 적용 범위를 규정할 때 "AI 시스템"이라는 용어를 사용하고, 사이버보안의 하위 집합으로 간주되는 "AI 보안"이라는 용어를 사용했다. 이 실행규범은 연구 목적으로만 AI 시스템을 생성하고 테스트하는 학계 연구자들(배포되지 않을

going to be deployed). The Code sets out cyber security requirements for the lifecycle of AI. We recognise that there is no consistent view in international frameworks on what forms the AI lifecycle. However, to help stakeholders, we have separated the principles into five phases. These are secure design, secure development, secure deployment, secure maintenance and secure end of life. We have also signposted relevant standards and publications at the start of each principle to highlight links between the various documents and the Code. This is not an exhaustive list.2	AI 시스템)을 대상으로 하지 않는다. 이 실행규범은 AI의 생명주기에 대한 사이버보안 요건을 제시한다. AI 생명주기가 무엇으로 구성되는지에 대한 국제 프레임워크들의 일관된 견해가 없다는 것을 인정한다. 그러나 이해관계자들을 돕기 위해, 원칙을 다섯 단계로 구분했다. 이는 안전한 설계, 안전한 개발, 안전한 배포, 안전한 유지관리, 안전한 생명주기 종료이다. 또한 각 원칙의 시작 부분에 관련 표준과 출판물을 안내하여 다양한 문서와 실행규범 간의 연결을 강조했다. 이는 모든 것을 망라한 목록은 아니다.
Implementation guide Following Call for Views feedback, we have created an implementation guide to support organisations with adhering to the requirements in the voluntary Code (and future standard). The guide was developed following an extensive review of software and AI standards and frameworks as well as documents published by other governments and regulators. The UK government plan to submit the Code and Implementation Guide in ETSI so that the future standard is accompanied by a guide. The government will update the content of the Code and Guide to mirror the future ETSI global standard and guide.	구현 가이드 의견 수렴 피드백에 따라, 조직이 자발적 실행규범(및 향후 표준)의 요건을 준수하는 것을 지원하기 위한 구현 가이드를 작성했다. 이 가이드는 소프트웨어 및 AI 표준과 프레임워크뿐만 아니라 다른 정부와 규제기관이 발표한 문서들에 대한 광범위한 검토를 통해 개발되었다. 영국 정부는 향후 표준이 가이드와 함께 제공되도록 ETSI에 실행규범과 구현 가이드를 제출할 계획이다. 정부는 향후 ETSI 글로벌 표준과 가이드를 반영하여 실행규범과 가이드의 내용을 업데이트할 것이다.
Audience This section defines the stakeholder groups that form the AI supply chain. An indication is given for each principle on which stakeholders are primarily responsible for its implementation. Importantly, a single entity may hold multiple stakeholder roles in this voluntary Code as well as responsibilities from different regulatory regimes.3 All stakeholders included in the table below should note that when the data used for an AI system is personal (including pseudonymized data), they may have data protection obligations and will need to consult UK data protection guidance offered by the ICO. Additionally, senior leaders in an organisation also have responsibilities to help protect their staff and infrastructure as noted in DSIT's Cyber Governance Code of Practice. Some provisions for Developers in the Code are less applicable to AI systems involving open-source models. We encourage Developers to review the Implementation Guide to confirm what requirements are specified for different types of AI systems.	대상 이 섹션은 AI 공급망을 구성하는 이해관계자 그룹을 정의한다. 각 원칙에 대해 어떤 이해관계자가 주로 구현 책임을 지는지 표시되어 있다. 중요한 것은, 단일 개체가 이 자발적 실행규범에서 여러 이해관계자 역할과 다른 규제 체계의 책임을 동시에 가질 수 있다는 것이다. 아래 표에 포함된 모든 이해관계자는 AI 시스템에 사용되는 데이터가 개인정보(가명화된 데이터 포함)인 경우, 데이터 보호 의무가 있을 수 있으며 ICO가 제공하는 영국 데이터 보호 가이드를 참조해야 한다는 점을 유의해야 한다. 또한, 조직의 고위 지도자들은 DSIT의 [사이버 거버넌스 실행규범]에 명시된 바와 같이 직원과 인프라를 보호하는 데 도움을 줄 책임이 있다. 실행규범의 개발자를 위한 일부 조항은 오픈소스 모델을 포함하는 AI 시스템에는 적용되지 않을 수 있다. 개발자들이 다양한 유형의 AI 시스템에 대해 명시된 요건을 확인하기 위해 구현 가이드를 검토하기를 권장한다.

Stakeholder	Definitions
Developers	This encompasses any type of business or organisation across any sector as well as

이해관계자	정의
개발자	AI 모델 및/또는 시스템을 생성하거나 적응시키는 책임을 지는 모든 유형의 사업체나 조직, 그리고 개인을 포함한다.

	individuals that are responsible for creating or adapting an AI model and/or system. This applies to all AI technologies, including proprietary and open-source models. For context, a business or organisation that creates an AI model and who is also responsible for embedding/deploying that model/system in their organisation would be defined in this voluntary Code to be both a Developer and a System Operator.
System Operators	This includes any type of business or organisation across any sector that has responsibility for embedding / deploying an AI model and system within their infrastructure. This applies to all AI technologies, including proprietary and open-source models. This term also includes those businesses that provide a contractual service to organisations to embed / deploy an AI model and system for business purposes.
Data Custodians	This includes any type of business, organisation or individual that controls data permissions and the integrity of data that is used for any AI model or system to function. This stakeholder group also includes those entities that set the policies for how data is used and managed for an AI model and/or system. In the context of an AI system, there could be multiple data custodians involved because some data used to create a model could come from the organisation that is deploying/embedding the system in their infrastructure and other data could be from public databases and other sources.
End-users	This encompasses any employee within an organisation or business and UK consumers who use an AI model and system for any purpose, including to support their work and day-to-day activities. This applies to all AI technologies and both proprietary and open-source models. This stakeholder group has been created because the voluntary Code has placed expectations on Developers, System Operators and Data custodians to help inform and protect end-users.
Affected entities	Encompasses all individuals and technologies, such as apps and autonomous systems, that are not directly affected by AI systems or decisions based on the output of AI systems. These individuals do not necessarily interact with the deployed system or application.

The table below gives examples of common cases involving different types of organisations that are relevant to this voluntary Code of Practice as well as the Software Resilience voluntary Code of Practice.

Stakeholder Groups	Guidance
Software vendors who	These organisations are likely to

	이는 독점 모델과 오픈소스 모델을 포함한 모든 AI 기술에 적용된다. 참고로, AI 모델을 생성하고 동시에 해당 모델/시스템을 자신의 조직에 임베드/배포하는 책임을 지는 사업체나 조직은 이 자발적 실행규범에서 개발자이자 시스템 운영자로 정의된다.
시스템 운영자	자신의 인프라 내에 AI 모델과 시스템을 임베드/배포하는 책임을 지는 모든 분야의 모든 유형의 사업체나 조직을 포함한다. 이는 독점 모델과 오픈소스 모델을 포함한 모든 AI 기술에 적용된다. 이 용어는 또한 사업 목적으로 조직을 위해 AI 모델과 시스템을 임베드/배포하는 계약 서비스를 제공하는 사업체들도 포함한다.
데이터 관리자	AI 모델이나 시스템이 기능하기 위해 사용되는 데이터의 권한과 무결성을 통제하는 모든 유형의 사업체, 조직 또는 개인을 포함한다. 이 이해관계자 그룹은 또한 AI 모델 및/또는 시스템에 대해 데이터가 사용되고 관리되는 방식에 대한 정책을 설정하는 개체들을 포함한다. AI 시스템의 맥락에서는 여러 데이터 관리자가 관련될 수 있다. 왜냐하면 모델을 생성하는 데 사용되는 일부 데이터는 시스템을 자신의 인프라에 배포/임베드하는 조직에서 올 수 있고, 다른 데이터는 공개 데이터베이스 및 기타 소스에서 올 수 있기 때문이다.
최종사용자	업무와 일상 활동을 지원하는 것을 포함하여 어떤 목적으로든 AI 모델과 시스템을 사용하는 조직이나 사업체 내의 모든 직원과 영국 소비자를 포함한다. 이는 독점 모델과 오픈소스 모델 모두를 포함한 모든 AI 기술에 적용된다. 이 이해관계자 그룹은 자발적 실행규범이 개발자, 시스템 운영자, 데이터 관리자에게 최종사용자를 알리고 보호하는 데 도움을 줄 것을 기대하고 있기 때문에 만들어졌다.
영향을 받는 개체	AI 시스템에 의해 직접적으로 영향을 받거나 AI 시스템의 출력에 기반한 결정에 의해 영향을 받는 모든 개인과 앱 및 자율 시스템과 같은 기술을 포함한다. 이러한 개인들은 배포된 시스템이나 애플리케이션과 반드시 상호작용하지는 않는다.

아래 표는 이 자발적 실천 강령과 소프트웨어 복원력 자발적 실천 강령과 관련된 다양한 유형의 조직이 관련된 일반적인 사례의 예를 보여준다.

이해관계자 그룹	가이드

also offer AI services to customers/end-users	be Developers and therefore are in scope of this Code and the Software Resilience Code of Practice.
Software vendors who use AI in their own infrastructure which has been created by an external provider	These organisations are likely to be System Operators and therefore are in scope of relevant parts of the Code and the Software Resilience Code of Practice.
Software vendors who create AI in-house and implement it within their infrastructure	These organisations are likely to be Developers and System Operators and therefore are in scope of this Code and the Software Resilience Code of Practice.
Software vendors who only use third-party AI (components) for their in-house use	These organisations are likely to be System Operators and therefore are in scope of relevant parts of the Code and the Software Resilience Code of Practice.
Organisation that creates an AI system for in-house use	These organisations are likely to be Developers and therefore are in scope of this Code.
Organisation that only uses third-party AI components	These organisations are likely to be System Operators and therefore are in scope of relevant parts of the Code.
AI Vendors	Organisations that offer or sell models and components, but do not play a role in developing or deploying them, are not likely to be in scope of this Code. These organisations are likely to be in scope of the Software Code of Practice and Cyber Governance Code.

고객/최종사용자에게 AI 서비스도 제공하는 소프트웨어 공급업체	이러한 조직은 개발자일 가능성이 높으며, 따라서 이 실행규범과 소프트웨어 복원력 실행규범의 적용 범위에 포함된다.
외부 제공업체가 생성한 AI를 자체 인프라에서 사용하는 소프트웨어 공급업체	이러한 조직은 시스템 운영자일 가능성이 높으며, 따라서 실행규범과 소프트웨어 복원력 실행규범의 관련 부분의 적용 범위에 포함된다.
사내에서 AI를 생성하고 자체 인프라 내에서 구현하는 소프트웨어 공급업체	이러한 조직은 개발자이자 시스템 운영자일 가능성이 높으며, 따라서 이 실행규범과 소프트웨어 복원력 실행규범의 적용 범위에 포함된다.
사내 사용을 위해 제3자 AI(구성 요소)만 사용하는 소프트웨어 공급업체	이러한 조직은 시스템 운영자일 가능성이 높으며, 따라서 실행규범과 소프트웨어 복원력 실행규범의 관련 부분의 적용 범위에 포함된다.
사내 사용을 위해 AI 시스템을 생성하는 조직	이러한 조직은 개발자일 가능성이 높으며, 따라서 이 실행규범의 적용 범위에 포함된다.
제3자 AI 구성 요소만 사용하는 조직	이러한 조직은 시스템 운영자일 가능성이 높으며, 따라서 실행규범의 관련 부분의 적용 범위에 포함된다.
AI 공급업체	모델과 구성 요소를 제공하거나 판매하지만 개발이나 배포에 역할을 하지 않는 조직은 이 실행규범의 적용 범위에 포함되지 않을 가능성이 높다. 이러한 조직은 소프트웨어 실행규범과 사이버 거버넌스 실행규범의 적용 범위에 포함될 가능성이 높다.

Terminology

We have used “shall” and “should” terminology for each provision in the voluntary Code to align with the wording used by standards development organisations. The table below sets out the definitions of these words in the context of the voluntary nature of this Code of Practice. A glossary can be found in Annex A.

Term	Definition
Shall	Indicates a requirement for the voluntary Code
Should	Indicates a recommendation for the voluntary Code

용어

표준 개발 조직에서 사용하는 용어에 맞추기 위해 자발적 실행규범의 각 조항에 대해 "shall"과 "should" 용어를 사용했다. 아래 표는 이 실행규범의 자발적 성격의 맥락에서 이러한 단어들의 정의를 제시한다. 용어집은 부록 A에서 찾을 수 있다.

용어	정의
Shall(해야 한다)	자발적 실행규범의 요건을 나타냄
Should (해야 할 것이다)	자발적 실행규범의 권고사항을 나타냄

May Indicates where something is possible, for example, that an organisation or individual is able to do something	May (할 수 있다) 어떤 것이 가능한 경우를 나타냄, 예를 들어 조직이나 개인이 어떤 것을 할 수 있음을 의미
Structure of the voluntary Code of Practice Principle 1: Raise awareness of AI security threats and risks Principle 2: Design your AI system for security as well as functionality and performance Principle 3: Evaluate the threats and manage the risks to your AI system Principle 4: Enable human responsibility for AI systems Principle 5: Identify, track and protect your assets Principle 6: Secure your infrastructure Principle 7: Secure your supply chain Principle 8: Document your data, models and prompts Principle 9: Conduct appropriate testing and evaluation Principle 10: Communication and processes associated with End-users and Affected Entities Principle 11: Maintain regular security updates, patches and mitigations Principle 12: Monitor your system's behaviour Principle 13: Ensure proper data and model disposal	자발적 실행규범의 구조 원칙 1: AI 보안 위협과 위험에 대한 인식 제고 원칙 2: 기능성과 성능뿐만 아니라 보안을 위한 AI 시스템 설계 원칙 3: 위협 평가 및 AI 시스템 위험 관리 원칙 4: AI 시스템에 대한 인간의 책임 활성화 원칙 5: 자산 식별, 추적 및 보호 원칙 6: 인프라 보안 원칙 7: 공급망 보안 원칙 8: 데이터, 모델 및 프롬프트 문서화 원칙 9: 적절한 테스트 및 평가 수행 원칙 10: 최종사용자 및 영향을 받는 개체와 관련된 커뮤니케이션 및 프로세스 원칙 11: 정기적인 보안 업데이트, 패치 및 완화 조치 유지 원칙 12: 시스템 행동 모니터링 원칙 13: 적절한 데이터 및 모델 폐기 보장
Code of Practice Principles Secure Design Principle 1: Raise awareness of AI security threats and risks Primarily applies to: System Operators, Developers, and Data Custodians [NIST 2022, NIST 2023, ASD 2023, WEF 2024, OWASP 2024, MITRE 2024, Google 2023, ESLA 2023, Cisco 2022, Deloitte 2023, Microsoft 2022]. 1.1. Organisations' cyber security training programme shall include AI security content which shall be regularly reviewed and updated, such as if new substantial AI-related security threats emerge. 1.1.1 AI security training shall be tailored to the specific roles and responsibilities of staff members. 1.2. As part of an Organisation's wider staff training programme, they shall require all staff to maintain awareness of the latest security threats and vulnerabilities that are AI-related. Where available, this awareness shall include	실행규범 원칙 안전한 설계 원칙 1: AI 보안 위협과 위험에 대한 인식 제고 주요 적용 대상: 시스템 운영자, 개발자, 데이터 관리자 [NIST 2022, NIST 2023, ASD 2023, WEF 2024, OWASP 2024, MITRE 2024, Google 2023, ESLA 2023, Cisco 2022, Deloitte 2023, Microsoft 2022]. 1.1 조직의 사이버보안 교육 프로그램은 AI 보안 콘텐츠를 포함해야 하며, 이는 새로운 중대한 AI 관련 보안 위협이 출현할 때와 같이 정기적으로 검토되고 업데이트되어야 한다. 1.1.1 AI 보안 교육은 직원의 특정 역할과 책임에 맞게 조정되어야 한다. 1.2 조직의 광범위한 직원 교육 프로그램의 일환으로, 모든 직원이 AI 관련 최신 보안 위협과 취약점에 대한 인식을 유지하도록 요구해야 한다. 이용 가능한 경우, 이러한

proposed mitigations.	인식에는 제안된 완화 조치가 포함되어야 한다.
1.2.1. These updates should be communicated through multiple channels, such as security bulletins, newsletters, or internal knowledge-sharing platforms. This will ensure broad dissemination and understanding among the staff.	1.2.1 이러한 업데이트는 보안 게시판, 뉴스레터 또는 내부 지식 공유 플랫폼과 같은 여러 채널을 통해 전달되어야 한다. 이는 직원들 사이에서 광범위한 배포와 이해를 보장할 것이다.
1.2.2 Organisations shall provide developers with training in secure coding and system design techniques specific to AI development, with a focus on preventing and mitigating security vulnerabilities in AI algorithms, models, and associated software.	1.2.2 조직은 개발자에게 AI 개발에 특화된 안전한 코딩 및 시스템 설계 기술에 대한 교육을 제공해야 하며, AI 알고리즘, 모델 및 관련 소프트웨어의 보안 취약점을 예방하고 완화하는 데 중점을 두어야 한다.
Principle 2: Design your AI system for security as well as functionality and performance Primarily applies to: System Operators and Developers [OWASP 2024, MITRE 2024, WEF 2024, ENISA 2023, NCSC 2023, BSI1 2023, Cisco 2022, Microsoft 2022, G7 2023, HHS 2021, OpenAI2 2024, ASD 2023, ICO 2020].	원칙 2: 기능성과 성능뿐만 아니라 보안을 위한 AI 시스템 설계 주요 적용 대상: 시스템 운영자 및 개발자 [OWASP 2024, MITRE 2024, WEF 2024, ENISA 2023, NCSC 2023, BSI1 2023, Cisco 2022, Microsoft 2022, G7 2023, HHS 2021, OpenAI2 2024, ASD 2023, ICO 2020].
2.1 As part of deciding whether to create an AI system, a System Operator and/or Developer shall conduct a thorough assessment that includes determining and documenting the business requirements and/or problem they are seeking to address, along with associated AI security risks and mitigation strategies.5	2.1 AI 시스템을 생성할지 결정하는 과정의 일환으로, 시스템 운영자 및/또는 개발자는 해결하고자 하는 사업적 요구사항 및/또는 문제와 관련 AI 보안 위험 및 완화 전략을 결정하고 문서화하는 것을 포함하는 철저한 평가를 수행해야 한다.
2.1.1 Where the Data Custodian is part of a Developer's organisation, they shall be included in internal discussions when determining the requirements and data needs of an AI system.	2.1.1 데이터 관리자가 개발자의 조직의 일부인 경우, AI 시스템의 요구사항과 데이터 필요를 결정할 때 내부 논의에 포함되어야 한다.
2.2: Developers and System Operators shall ensure that AI systems are designed and implemented to withstand adversarial AI attacks, unexpected inputs and AI system failure.	2.2 개발자와 시스템 운영자는 AI 시스템이 적대적 AI 공격, 예상치 못한 입력 및 AI 시스템 장애에 견딜 수 있도록 설계되고 구현되어야 한다.
2.3 To support the process of preparing data, security auditing and incident response for an AI system, Developers shall document and create an audit trail in relation to the AI system. This shall include the operation, and lifecycle management of models, datasets and prompts incorporated into the system.	2.3 AI 시스템의 데이터 준비, 보안 감사 및 사고 대응 과정을 지원하기 위해, 개발자는 AI 시스템과 관련하여 문서화하고 감사 추적을 생성해야 한다. 이는 시스템에 통합된 모델, 데이터세트 및 프롬프트의 운영과 생명주기 관리를 포함해야 한다.

2.4 If a Developer or System Operator uses an external component, they shall conduct an AI security risk assessment and due diligence process in line with their existing software development processes, that assesses AI specific risks.6	2.4 개발자나 시스템 운영자가 외부 구성 요소를 사용하는 경우, AI 특화 위험을 평가하는 기존 소프트웨어 개발 프로세스에 따라 AI 보안 위험 평가와 실사 과정을 수행해야 한다.
2.5 Data Custodians shall ensure that the intended usage of the system is appropriate to the sensitivity of the data it was trained on as well as the controls intended to ensure the security of the data.	2.5 데이터 관리자는 시스템의 의도된 사용이 훈련에 사용된 데이터의 민감성과 데이터 보안을 보장하기 위한 통제 조치에 적합한지 확인해야 한다.
2.5.1 Organisations should ensure that employees are encouraged to proactively report and identify any potential security risks in AI systems and ensure appropriate safeguards are in place.	2.5.1 조직은 직원들이 AI 시스템의 잠재적 보안 위험을 적극적으로 보고하고 식별하도록 장려하고 적절한 보호 조치가 마련되어 있는지 확인해야 한다.
2.6 Where the AI system will be interacting with other systems or data sources, (be they internal or external), Developers and System Operators shall ensure that the permissions granted to the AI system on other systems are only provided as required for functionality and are risk assessed.	2.6 AI 시스템이 다른 시스템이나 데이터 소스(내부 또는 외부)와 상호작용하는 경우, 개발자와 시스템 운영자는 AI 시스템에 부여된 다른 시스템에 대한 권한이 기능에 필요한 경우에만 제공되고 위험 평가가 수행되어야 한다.
2.7 If a Developer or System Operator chooses to work with an external provider, they shall undertake a due diligence assessment and should ensure that the provider is adhering to this Code of Practice.	2.7 개발자나 시스템 운영자가 외부 제공업체와 협력하기로 선택하는 경우, 실사 평가를 수행해야 하며 제공업체가 이 실행규범을 준수하고 있는지 확인해야 한다.
Principle 3: Evaluate the threats and manage the risks to your AI system Primarily applies to: Developers and System Operators [OWASP 2024, WEF 2024, Nvidia 2023, ENISA 2023, Google 2023, G7 2023, NCSC 2023, Deloitte 2023], MITRE, OWASP, NIST Risk Taxonomy, ISO 27001]	원칙 3: 위협 평가 및 AI 시스템 위험 관리 주요 적용 대상: 개발자 및 시스템 운영자 [OWASP 2024, WEF 2024, Nvidia 2023, ENISA 2023, Google 2023, G7 2023, NCSC 2023, Deloitte 2023], MITRE, OWASP, NIST Risk Taxonomy, ISO 27001]
3.1 Developers and System Operators shall analyse threats and manage security risks to their systems. Threat modelling should include regular reviews and updates and address AI-specific attacks, such as data poisoning, model inversion, and membership inference.	3.1 개발자와 시스템 운영자는 시스템에 대한 위협을 분석하고 보안 위험을 관리해야 한다. 위협 모델링은 정기적인 검토와 업데이트를 포함해야 하며 데이터 오염, 모델 역전, 멤버십 추론과 같은 AI 특화 공격을 다뤄야 한다.
3.1.1 The threat modelling and risk management process shall be conducted to address any security risks that arise when a new setting or configuration option is implemented or updated at any stage of the AI lifecycle.	3.1.1 위협 모델링과 위험 관리 과정은 AI 생명주기의 어느 단계에서든 새로운 설정이나 구성 옵션이 구현되거나 업데이트될 때 발생하는 모든 보안 위험을 해결하기 위해 수행되어야 한다.

3.1.2 Developers shall manage the security risks associated with AI models that provide superfluous functionalities, where increased functionality leads to increased risk. For example, where a multi-modal model is being used but only single modality is used for system function.	3.1.2 개발자는 불필요한 기능을 제공하는 AI 모델과 관련된 보안 위험을 관리해야 한다. 여기서 증가된 기능성은 증가된 위험으로 이어진다. 예를 들어, 다중 모달 모델이 사용되지만 시스템 기능에는 단일 모달리티만 사용되는 경우이다.
3.1.3 System Operators shall apply controls to risks identified through the analysis based on a range of considerations, including the cost of implementation in line with their corporate risk tolerance.	3.1.3 시스템 운영자는 기업의 위험 허용도에 따른 구현 비용을 포함한 다양한 고려사항을 바탕으로 분석을 통해 식별된 위험에 대한 통제를 적용해야 한다.
3.2 Where AI security threats are identified that cannot be resolved by Developers, this shall be communicated to System Operators so they can threat model their systems. System Operators shall communicate this information to End-users, so they are made aware of these threats. This communication should include detailed descriptions of the risks, potential impacts, and recommended actions to address or monitor these threats.	3.2 개발자가 해결할 수 없는 AI 보안 위협이 식별되는 경우, 시스템 운영자가 자신의 시스템에 대한 위협 모델링을 할 수 있도록 이를 전달해야 한다. 시스템 운영자는 최종사용자가 이러한 위협을 알 수 있도록 이 정보를 전달해야 한다. 이러한 커뮤니케이션에는 위험의 세부 설명, 잠재적 영향, 이러한 위협을 해결하거나 모니터링하기 위한 권장 조치가 포함되어야 한다.
3.3 Where an external entity has responsibility for AI security risks identified within an organisations infrastructure, System Operators should attain assurance that these parties are able to address such risks.	3.3 외부 개체가 조직의 인프라 내에서 식별된 AI 보안 위험에 대한 책임을 지는 경우, 시스템 운영자는 이러한 당사자들이 그러한 위험을 해결할 수 있다는 보장을 얻어야 한다.
3.4 Developers and System Operators should continuously monitor and review their system infrastructure according to risk appetite. It is important to recognise that a higher level of risk will remain in AI systems despite the application of controls to mitigate against them.	3.4 개발자와 시스템 운영자는 위험 선호도에 따라 시스템 인프라를 지속적으로 모니터링하고 검토해야 한다. 통제 조치를 적용하여 완화하더라도 AI 시스템에는 더 높은 수준의 위험이 남아있을 것임을 인식하는 것이 중요하다.
Principle 4: Enable human responsibility for AI systems Primarily applies to: Developers and System Operators	원칙 4: AI 시스템에 대한 인간의 책임 활성화 주요 적용 대상: 개발자 및 시스템 운영자
[OWASP 2024, MITRE 2024, BSI1 2023, Microsoft 2022]	[OWASP 2024, MITRE 2024, BSI1 2023, Microsoft 2022]
4.1 When designing an AI system, Developers and/or System Operators should incorporate and maintain capabilities to enable human oversight.7	4.1 AI 시스템을 설계할 때, 개발자 및/또는 시스템 운영자는 인간의 감독을 가능하게 하는 기능을 통합하고 유지해야 한다.
4.2 Developers should design systems to make it easy for humans to assess outputs that they are responsible for in said system (such as by ensuring that models outputs are explainable or interpretable).	4.2 개발자는 인간이 해당 시스템에서 자신이 책임지는 출력을 평가하기 쉽도록 시스템을 설계해야 한다(예: 모델 출력이 설명 가능하거나 해석 가능하도록 보장).

4.3 Where human oversight is a risk control, Developers and/or System Operators shall design, develop, verify and maintain technical measures to reduce the risk through such oversight.	4.3 인간의 감독이 위험 통제인 경우, 개발자 및/또는 시스템 운영자는 그러한 감독을 통해 위험을 줄이는 기술적 조치를 설계, 개발, 검증 및 유지해야 한다.
4.4 Developers should verify that the security controls specified by the Data Custodian have been built into the system.	4.4 개발자는 데이터 관리자가 명시한 보안 통제가 시스템에 구축되었는지 검증해야 한다.
4.5 Developers and System Operators should make End-users aware of prohibited use cases of the AI system.	4.5 개발자와 시스템 운영자는 최종사용자에게 AI 시스템의 금지된 사용 사례를 알려야 한다.
Secure Development Principle 5: Identify, track and protect your assets Primarily applies to: Developers, System Operators and Data Custodians [OWASP 2024, Nvidia 2023, NCSC 2023, BSI1 2023, Cisco 2022, Deloitte 2023, Amazon 2023, G7 2023, ICO 2020]	원칙 5: 자산 식별, 추적 및 보호 주요 적용 대상: 개발자, 시스템 운영자 및 데이터 관리자 [OWASP 2024, Nvidia 2023, NCSC 2023, BSI1 2023, Cisco 2022, Deloitte 2023, Amazon 2023, G7 2023, ICO 2020]
5.1 Developers, Data Custodians and System Operators shall maintain a comprehensive inventory of their assets (including their interdependencies/connectivity).	5.1 개발자, 데이터 관리자 및 시스템 운영자는 자산의 포괄적인 목록(상호 의존성/연결성 포함)을 유지해야 한다.
5.2 As part of broader software security practices, Developers, Data Custodians and System Operators shall have processes and tools to track, authenticate, manage version control and secure their assets due to the increased complexities of AI specific assets.	5.2 광범위한 소프트웨어 보안 관행의 일환으로, 개발자, 데이터 관리자 및 시스템 운영자는 AI 특화 자산의 증가된 복잡성으로 인해 자산을 추적, 인증, 버전 관리 및 보안하는 프로세스와 도구를 갖춰야 한다.
5.3 System Operators shall develop and tailor their disaster recovery plans to account for specific attacks aimed at AI systems.	5.3 시스템 운영자는 AI 시스템을 겨냥한 특정 공격을 고려하여 재해 복구 계획을 개발하고 맞춤화해야 한다.
5.3.1 System Operators should ensure that a known good state can be restored.	5.3.1 시스템 운영자는 알려진 정상 상태가 복원될 수 있도록 보장해야 한다.
5.4 Developers, System Operators, Data Custodians and End-users shall protect sensitive data, such as training or test data, against unauthorised access.	5.4 개발자, 시스템 운영자, 데이터 관리자 및 최종사용자는 훈련 또는 테스트 데이터와 같은 민감한 데이터를 무단 접근으로부터 보호해야 한다.
5.4.1 Developers, Data Custodians and System Operators shall apply checks and sanitisation to data and inputs when designing the model based on their access to said data and inputs and where those data and inputs are stored. This shall be repeated when model revisions are made in response to user feedback or continuous learning.	5.4.1 개발자, 데이터 관리자 및 시스템 운영자는 해당 데이터와 입력에 대한 접근과 그러한 데이터와 입력이 저장된 위치를 기반으로 모델을 설계할 때 데이터와 입력에 대한 검사와 정제를 적용해야 한다. 이는 사용자 피드백이나 지속적 학습에 대응하여 모델 수정이 이루어질 때

	반복되어야 한다.
5.4.2 Where training data or model weights could be confidential, Developers shall put proportionate protections in place.	5.4.2 훈련 데이터나 모델 가중치가 기밀일 수 있는 경우, 개발자는 비례적인 보호 조치를 마련해야 한다.
Principle 6: Secure your infrastructure Primarily applies to: Developers and System Operators [OWASP 2024, MITRE 2024, WEF 2024, NCSC 2023, Microsoft 2022, ICO 2020]	원칙 6: 인프라 보안 주요 적용 대상: 개발자 및 시스템 운영자 [OWASP 2024, MITRE 2024, WEF 2024, NCSC 2023, Microsoft 2022, ICO 2020]
6.1 Developers and System Operators shall evaluate their organisation's access control frameworks and identify appropriate measures to secure APIs, models, data, and training and processing pipelines.	6.1 개발자와 시스템 운영자는 조직의 접근 통제 프레임워크를 평가하고 API, 모델, 데이터, 훈련 및 처리 파이프라인을 보안하기 위한 적절한 조치를 식별해야 한다.
6.2 If a Developer offers an API to external customers or collaborators, they shall apply controls that mitigate attacks on the AI system via the API. For example, placing limits on model access rate to limit an attacker's ability to reverse engineer or overwhelm defences to rapidly poison a model.	6.2 개발자가 외부 고객이나 협력자에게 API를 제공하는 경우, API를 통한 AI 시스템 공격을 완화하는 통제를 적용해야 한다. 예를 들어, 공격자가 역공학하거나 방어를 압도하여 모델을 빠르게 오염시키는 능력을 제한하기 위해 모델 접근 속도에 제한을 두는 것이다.
6.3 Developers shall also create dedicated environments for development and model tuning activities. The dedicated environments shall be backed by technical controls to ensure separation and principle of least privilege. In the context of AI, this is particularly necessary because training data shall only be present in the training and development environments where this training data is not based on publicly available data.	6.3 개발자는 또한 개발 및 모델 튜닝 활동을 위한 전용 환경을 만들어야 한다. 전용 환경은 분리와 최소 권한 원칙을 보장하는 기술적 통제에 의해 뒷받침되어야 한다. AI의 맥락에서, 이는 특히 필요하다. 왜냐하면 훈련 데이터가 공개적으로 이용 가능한 데이터를 기반으로 하지 않는 경우 훈련 데이터는 훈련 및 개발 환경에만 존재해야 하기 때문이다.
6.4 Developers and System Operators shall implement and publish a clear and accessible vulnerability disclosure policy.	6.4 개발자와 시스템 운영자는 명확하고 접근 가능한 취약점 공개 정책을 구현하고 공개해야 한다.
6.5 Developers and System Operators shall create, test and maintain an AI system incident management plan and an AI system recovery plan.	6.5 개발자와 시스템 운영자는 AI 시스템 사고 관리 계획과 AI 시스템 복구 계획을 생성, 테스트 및 유지해야 한다.
6.6 Developers and System Operators should ensure that, where they are using cloud service operators to help to deliver the capability, their contractual agreements support compliance with the above requirements.	6.6 개발자와 시스템 운영자는 능력 제공을 돕기 위해 클라우드 서비스 운영자를 사용하는 경우, 계약 협정이 상기 요구사항의 준수를 지원하는지 확인해야 한다.
Principle 7: Secure your supply chain Primarily applies to: Developers, System Operators and Data	원칙 7: 공급망 보안 주요 적용 대상: 개발자, 시스템 운영자 및 데이터 관리자

Custodians [Software Bill of Materials (SBOM), CISA, OWASP 2024, NCSC 2023, Microsoft 2022, ASD 2023]	[Software Bill of Materials (SBOM), CISA, OWASP 2024, NCSC 2023, Microsoft 2022, ASD 2023]
7.1 Developers and System Operators shall follow secure software supply chain processes for their AI model and system development.	7.1 개발자와 시스템 운영자는 AI 모델 및 시스템 개발을 위해 안전한 소프트웨어 공급망 프로세스를 따라야 한다.
7.2 System Operators that choose to use or adapt any models, or components, which are not well-documented or secured shall be able to justify their decision to use such models or components through documentation (for example if there was no other supplier for said component).	7.2 잘 문서화되지 않았거나 보안이 되지 않은 모델이나 구성 요소를 사용하거나 적응시키기로 선택한 시스템 운영자는 문서화를 통해 그러한 모델이나 구성 요소를 사용하기로 한 결정을 정당화할 수 있어야 한다(예: 해당 구성 요소에 대한 다른 공급업체가 없었던 경우).
7.2.1 In this case, Developers and System Operators shall have mitigating controls and undertake a risk assessment linked to such models or components.	7.2.1 이 경우, 개발자와 시스템 운영자는 완화 통제를 갖추고 그러한 모델이나 구성 요소와 연결된 위험 평가를 수행해야 한다.
7.2.2 System Operators shall share this documentation with End-users in an accessible way.	7.2.2 시스템 운영자는 이 문서를 최종사용자가 접근 가능한 방식으로 공유해야 한다.
7.3 Developers and System Operators shall re-run evaluations on released models that they intend on using.	7.3 개발자와 시스템 운영자는 사용하려는 출시된 모델에 대해 평가를 재실행해야 한다.
7.4 System Operators shall communicate their intention to update models to End-users in an accessible way prior to models being updated.	7.4 시스템 운영자는 모델이 업데이트되기 전에 모델을 업데이트하려는 의도를 최종사용자에게 접근 가능한 방식으로 전달해야 한다.
Principle 8: Document your data, models and prompts Primarily applies to: Developers [OWASP 2024, WEF 2024, NCSC 2023, Cisco 2022, Microsoft 2022, ICO 2020]	원칙 8: 데이터, 모델 및 프롬프트 문서화 주요 적용 대상: 개발자 [OWASP 2024, WEF 2024, NCSC 2023, Cisco 2022, Microsoft 2022, ICO 2020]
8.1 Developers shall document and maintain a clear audit trail of their system design and post-deployment maintenance plans. Developers should make the documentation available to the downstream System Operators and Data Custodians.	8.1 개발자는 시스템 설계와 배포 후 유지보수 계획의 명확한 감사 추적을 문서화하고 유지해야 한다. 개발자는 문서를 하위 시스템 운영자와 데이터 관리자가 이용할 수 있도록 해야 한다.
8.1.1 Developers should ensure that the document includes security-relevant information, such as the sources of training data (including fine-tuning data and human or other operational feedback), intended scope and limitations, guardrails, retention time, suggested review frequency and potential failure modes.	8.1.1 개발자는 문서에 훈련 데이터의 출처(미세 조정 데이터와 인간 또는 기타 운영 피드백 포함), 의도된 범위와 제한사항, 가드레일, 보유 시간, 제안된 검토 빈도 및

8.1.2 Developers shall release cryptographic hashes for model components that are made available to other stakeholders to allow them to verify the authenticity of the components. 8.2 Where training data has been sourced from publicly available sources, there is a risk that this data might have been poisoned. As discovery of poisoned data is likely to occur after training (if at all), Developers shall document how they obtained the public training data, where it came from and how that data is used in the model. 8.2.1. The documentation of training data should include at a minimum the source of the data, such as the URL of the scraped page, and the date/time the data was obtained. This will allow Developers to identify whether a reported data poisoning attack was in their data sets. 8.3 Developers should ensure that they have an audit log of changes to system prompts or other model configuration (including prompts) that affect the underlying working of the systems. Developers may make this available to any System Operators and End-Users that have access to the model.	잠재적 실패 모드와 같은 보안 관련 정보가 포함되도록 해야 한다. 8.1.2 개발자는 다른 이해관계자가 이용할 수 있게 된 모델 구성 요소에 대해 구성 요소의 진위성을 검증할 수 있도록 암호화 해시를 공개해야 한다. 8.2 훈련 데이터가 공개적으로 이용 가능한 소스에서 조달된 경우, 이 데이터가 오염되었을 위험이 있다. 오염된 데이터의 발견은 훈련 후에(또는 전혀) 발생할 가능성이 높으므로, 개발자는 공개 훈련 데이터를 어떻게 얻었는지, 어디서 왔는지, 그 데이터가 모델에서 어떻게 사용되는지 문서화해야 한다. 8.2.1 훈련 데이터의 문서화는 최소한 스크랩된 페이지의 URL과 같은 데이터 출처와 데이터가 얻어진 날짜/시간을 포함해야 한다. 이를 통해 개발자는 보고된 데이터 오염 공격이 자신의 데이터세트에 있었는지 식별할 수 있다. 8.3 개발자는 시스템의 기본 작동에 영향을 미치는 시스템 프롬프트나 기타 모델 구성(프롬프트 포함)의 변경사항에 대한 감사 로그를 갖춰야 한다. 개발자는 이를 모델에 접근할 수 있는 모든 시스템 운영자와 최종사용자가 이용할 수 있도록 할 수 있다.
Principle 9: Conduct appropriate testing and evaluation Primarily applies to: Developers and System Operators [OWASP 2024, WEF 2024, Nvidia 2023, NCSC 2023, ENISA 2023, Google 2023, G7 2023] 9.1 Developers shall ensure that all models, applications and systems that are released to System Operators and/or End-users have been tested as part of a security assessment process. 9.2 System Operators shall conduct testing prior to the system being deployed with support from Developers. 9.2.1 For security testing, System Operators and Developers should use independent security testers with technical skills relevant to their AI systems. 9.3 Developers should ensure that the findings from the	원칙 9: 적절한 테스트 및 평가 수행 주요 적용 대상: 개발자 및 시스템 운영자 [OWASP 2024, WEF 2024, Nvidia 2023, NCSC 2023, ENISA 2023, Google 2023, G7 2023] 9.1 개발자는 시스템 운영자 및/또는 최종사용자에게 출시되는 모든 모델, 애플리케이션 및 시스템이 보안 평가 과정의 일환으로 테스트되었는지 확인해야 한다. 9.2 시스템 운영자는 개발자의 지원을 받아 시스템이 배포되기 전에 테스트를 수행해야 한다. 9.2.1 보안 테스트의 경우, 시스템 운영자와 개발자는 자신의 AI 시스템과 관련된 기술적 기술을 가진 독립적인 보안 테스터를 사용해야 한다.

testing and evaluation are shared with System Operators, to inform their own testing and evaluation. 9.4 Developers should evaluate model outputs to ensure they do not allow System Operators or End-users to reverse engineer non-public aspects of the model or the training data. 9.4.1 Additionally, Developers should evaluate model outputs to ensure they do not provide System Operators or End-users with unintended influence over the system.	9.3 개발자는 테스트와 평가의 결과를 시스템 운영자와 공유하여 자신의 테스트와 평가에 정보를 제공해야 한다. 9.4 개발자는 모델 출력이 시스템 운영자나 최종사용자가 모델의 비공개 측면이나 훈련 데이터를 역공학하는 것을 허용하지 않는지 평가해야 한다. 9.4.1 또한, 개발자는 모델 출력이 시스템 운영자나 최종사용자에게 시스템에 대한 의도하지 않은 영향력을 제공하지 않는지 평가해야 한다.
Secure Deployment Principle 10: Communication and processes associated with End-users and Affected Entities As part of an organisation's wider deployment practices, they should also consider pre-deployment testing of AI systems alongside the requirements below. 10.1 System Operators shall convey to End-users in an accessible way where and how their data will be used, accessed and stored (for example, if it is used for model retraining, or reviewed by employees or partners).8 If the Developer is an external entity, they shall provide this information to System Operators. 10.2 System Operators shall provide End-users with accessible guidance to support their use, management, integration, and configuration of AI systems. If the Developer is an external entity, they shall provide all necessary information to help System Operators. 10.2.1 System Operators shall include guidance on the appropriate use of the model or system, which includes highlighting limitations and potential failure modes. 10.2.2 System Operators shall proactively inform End-users of any security relevant updates and provide clear explanations in an accessible way. 10.3 Developers and System Operators should support End-users and Affected Entities during and following a cyber security incident to contain and mitigate the impacts of an incident. The process for undertaking this should be documented and agreed in contracts with End-users.	안전한 배포 원칙 10: 최종사용자 및 영향을 받는 개체와 관련된 커뮤니케이션 및 프로세스 조직의 광범위한 배포 관행의 일환으로, 아래 요구사항과 함께 AI 시스템의 배포 전 테스트도 고려해야 한다. 10.1 시스템 운영자는 최종사용자의 데이터가 어디서 어떻게 사용되고, 접근되고, 저장될지(예: 모델 재훈련에 사용되거나 직원이나 파트너가 검토하는 경우)를 접근 가능한 방식으로 전달해야 한다. 개발자가 외부 개체인 경우, 시스템 운영자에게 이 정보를 제공해야 한다. 10.2 시스템 운영자는 최종사용자가 AI 시스템을 사용, 관리, 통합 및 구성하는 것을 지원하기 위한 접근 가능한 가이드를 제공해야 한다. 개발자가 외부 개체인 경우, 시스템 운영자를 돕기 위해 필요한 모든 정보를 제공해야 한다. 10.2.1 시스템 운영자는 제한사항과 잠재적 실패 모드를 강조하는 것을 포함하여 모델이나 시스템의 적절한 사용에 대한 가이드를 포함해야 한다. 10.2.2 시스템 운영자는 보안 관련 업데이트를 최종사용자에게 적극적으로 알리고 접근 가능한 방식으로 명확한 설명을 제공해야 한다. 10.3 개발자와 시스템 운영자는 사이버보안 사고 동안과 이후에 최종사용자와 영향을 받는 개체를 지원하여 사고의 영향을 억제하고 완화해야 한다. 이를 수행하는 과정은 문서화되고 최종사용자와의 계약에서 합의되어야 한다.
Secure Maintenance	원칙 11: 정기적인 보안 업데이트, 패치 및 완화 조치 유지

Principle 11: Maintain regular security updates, patches and mitigations Primarily applies to: Developers and System Operators [ICO 2020]	주요 적용 대상: 개발자 및 시스템 운영자 [ICO 2020]
11.1 Developers shall provide security updates and patches, where possible, and notify System Operators of the security updates. System Operators shall deliver these updates and patches to End-users.	11.1 개발자는 가능한 경우 보안 업데이트와 패치를 제공하고 시스템 운영자에게 보안 업데이트를 알려야 한다. 시스템 운영자는 이러한 업데이트와 패치를 최종사용자에게 전달해야 한다.
11.1.1 Developers shall have mechanisms and contingency plans to mitigate security risks, particularly in instances where updates cannot be provided for AI systems.	11.1.1 개발자는 특히 AI 시스템에 대해 업데이트를 제공할 수 없는 경우에 보안 위험을 완화하기 위한 메커니즘과 비상 계획을 갖춰야 한다.
11.2 Developers should treat major AI system updates as though a new version of a model has been developed and therefore undertake a new security testing and evaluation process to help protect users.	11.2 개발자는 주요 AI 시스템 업데이트를 새로운 버전의 모델이 개발된 것처럼 취급하고 따라서 사용자를 보호하는 데 도움이 되는 새로운 보안 테스트와 평가 과정을 수행해야 한다.
11.3 Developers should support System Operators to evaluate and respond to model changes, (for example by providing preview access via beta-testing and versioned APIs).	11.3 개발자는 시스템 운영자가 모델 변경을 평가하고 대응할 수 있도록 지원해야 한다(예: 베타 테스트와 버전화된 API를 통한 미리보기 접근 제공).
Principle 12: Monitor your system's behaviour Primarily applies to: Developers and System Operators	원칙 12: 시스템 행동 모니터링 주요 적용 대상: 개발자 및 시스템 운영자
[OWASP 2024, WEF 2024, Nvidia 2023, ENISA 2023, BSI1 2023, Cisco 2022, Deloitte 2023, G7 2023, Amazon 2023, ICO 2020]	[OWASP 2024, WEF 2024, Nvidia 2023, ENISA 2023, BSI1 2023, Cisco 2022, Deloitte 2023, G7 2023, Amazon 2023, ICO 2020]
12.1 System Operators shall log system and user actions to support security compliance, incident investigations, and vulnerability remediation.	12.1 시스템 운영자는 보안 준수, 사고 조사 및 취약점 수정을 지원하기 위해 시스템과 사용자 행동을 기록해야 한다.
12.2 System Operators should analyse their logs to ensure that AI models continue to produce desired outputs and to detect anomalies, security breaches, or unexpected behaviour over time (such as due to data drift or data poisoning).	12.2 시스템 운영자는 AI 모델이 계속해서 원하는 출력을 생성하는지 확인하고 시간이 지남에 따라(데이터 드리프트나 데이터 오염 등으로 인한) 이상 징후, 보안 침해 또는 예상치 못한 행동을 탐지하기 위해 로그를 분석해야 한다.
12.3 System Operators and Developers should monitor internal states of their AI systems where this could better enable them to address security threats, or to enable future security analytics.	12.3 시스템 운영자와 개발자는 보안 위협을 더 잘 해결하거나 향후 보안 분석을 가능하게 할 수 있는 경우 AI 시스템의 내부 상태를 모니터링해야 한다.
12.4 System Operators and Developers should monitor the	12.4 시스템 운영자와 개발자는 보안에 영향을 줄 수 있는 행동의 급격하거나 점진적인 변화를 탐지할 수 있도록

performance of their models and system over time so that they can detect sudden or gradual changes in behaviour that could affect security.	시간이 지남에 따라 모델과 시스템의 성능을 모니터링해야 한다.
Secure End of Life Principle 13: Ensure proper data and model disposal Primarily applies to: Developers and System Operators 13.1 If a Developer or System Operator decides to transfer or share ownership of training data and/or a model to another entity they shall involve Data Custodians and securely dispose of these assets. This will protect AI security issues that may transfer from one AI system instantiation to another. 13.2 If a Developer or System Operators decides to decommission a model and/or system, they shall involve Data Custodians and securely delete applicable data and configuration details.	안전한 생명주기 종료 원칙 13: 적절한 데이터 및 모델 폐기 보장 주요 적용 대상: 개발자 및 시스템 운영자 13.1 개발자나 시스템 운영자가 훈련 데이터 및/또는 모델의 소유권을 다른 개체에 이전하거나 공유하기로 결정하는 경우, 데이터 관리자를 포함시키고 이러한 자산을 안전하게 폐기해야 한다. 이는 한 AI 시스템 인스턴스에서 다른 시스템으로 이전될 수 있는 AI 보안 문제를 보호할 것이다. 13.2 개발자나 시스템 운영자가 모델 및/또는 시스템을 폐기하기로 결정하는 경우, 데이터 관리자를 포함시키고 해당하는 데이터와 구성 세부사항을 안전하게 삭제해야 한다.
Annex A: Glossary of terms Adversarial AI: Describes techniques and methods that exploit vulnerabilities in the way AI systems work, for example, by introducing malicious inputs to exploit their machine learning aspect and deceive the system into producing incorrect or unintended results. These techniques are commonly used in adversarial attacks but are not a distinct type of AI system. Adversarial Attack: An attempt to manipulate an AI model by introducing specially crafted inputs to cause the model to produce errors or unintended outcomes. Application Programming Interface (API): A set of tools and protocols that allow different software systems to communicate and interact. Artificial Intelligence (AI): Systems designed to perform tasks typically requiring human intelligence, such as decision-making, language understanding and pattern recognition. These systems can operate with varying levels of autonomy and adapt to their environment or data to improve performance. Data Poisoning: A type of adversarial attack where malicious data is introduced into training datasets to compromise the AI system's performance or behaviour. Explainability: The ability of an AI system to provide	부록 A: 용어집 적대적 AI: AI 시스템이 작동하는 방식의 취약점을 악용하는 기술과 방법을 설명한다. 예를 들어, 머신러닝 측면을 악용하고 시스템을 속여 잘못되거나 의도하지 않은 결과를 생성하도록 악의적인 입력을 도입하는 것이다. 이러한 기술은 일반적으로 적대적 공격에 사용되지만 별개의 AI 시스템 유형은 아니다. 적대적 공격: 모델이 오류나 의도하지 않은 결과를 생성하도록 특별히 제작된 입력을 도입하여 AI 모델을 조작하려는 시도이다. 애플리케이션 프로그래밍 인터페이스(API): 서로 다른 소프트웨어 시스템이 통신하고 상호작용할 수 있게 하는 도구와 프로토콜의 집합이다. 인공지능(AI): 의사결정, 언어 이해, 패턴 인식과 같이 일반적으로 인간의 지능이 필요한 작업을 수행하도록 설계된 시스템이다. 이러한 시스템은 다양한 수준의 자율성으로 작동할 수 있으며 성능을 향상시키기 위해 환경이나 데이터에 적응할 수 있다. 데이터 오염: AI 시스템의 성능이나 행동을 손상시키기 위해 훈련 데이터세트에 악의적인 데이터를 도입하는 적대적 공격의 한 유형이다. 설명 가능성: AI 시스템이 자신의 의사결정 과정에 대해

human-understandable insights into its decision-making process.	인간이 이해할 수 있는 통찰력을 제공하는 능력이다.
Guardrails: Predefined constraints or rules implemented to control and limit an AI system's outputs and behaviours, ensuring safety, reliability, and alignment with ethical or operational guidelines.	가드레일: 안전성, 신뢰성 및 윤리적 또는 운영적 가이드라인과의 정렬을 보장하기 위해 AI 시스템의 출력과 행동을 통제하고 제한하기 위해 구현된 사전 정의된 제약이나 규칙이다.
Inference Attack: A privacy attack where an adversary retrieves sensitive information about the training data, or users, by analysing the outputs of an AI model.	추론 공격: 적대자가 AI 모델의 출력을 분석하여 훈련 데이터나 사용자에 대한 민감한 정보를 검색하는 프라이버시 공격이다.
Model Inversion: A privacy attack where an adversary infers sensitive information about the training data by analysing the AI model's outputs.	모델 역전: 적대자가 AI 모델의 출력을 분석하여 훈련 데이터에 대한 민감한 정보를 추론하는 프라이버시 공격이다.
Prompt: An input provided to an AI model, often in the form of text, that directs or guides its response. Prompts can include questions, instructions, or context for the desired output.	프롬프트: 종종 텍스트 형태로 AI 모델에 제공되는 입력으로, 모델의 응답을 지시하거나 안내한다. 프롬프트에는 원하는 출력에 대한 질문, 지시사항 또는 맥락이 포함될 수 있다.
Risk Assessment: The process of identifying, analysing and mitigating potential threats to the security or functionality of an AI system.	위험 평가: AI 시스템의 보안이나 기능에 대한 잠재적 위협을 식별, 분석 및 완화하는 과정이다.
Sanitisation: The process of cleaning and validating data or inputs to remove errors, inconsistencies and malicious content, ensuring data integrity and security.	정제: 오류, 불일치 및 악의적인 콘텐츠를 제거하기 위해 데이터나 입력을 정리하고 검증하는 과정으로, 데이터 무결성과 보안을 보장한다.
System Prompt: A predefined input or set of instructions provided to guide the behaviour of an AI model, often used to define its tone, rules, or operational context.	시스템 프롬프트: AI 모델의 행동을 안내하기 위해 제공되는 사전 정의된 입력이나 지시사항 집합으로, 종종 모델의 톤, 규칙 또는 운영 맥락을 정의하는 데 사용된다.
Threat Modelling: A process to identify and address potential security threats to a system during its design and development phases.	위협 모델링: 시스템의 설계와 개발 단계에서 잠재적 보안 위협을 식별하고 해결하는 과정이다.
Training: The process of teaching an AI model to recognise patterns, make decisions, or generate outputs by exposing it to labelled data and adjusting its parameters to minimise errors.	훈련: 라벨이 지정된 데이터에 노출시키고 오류를 최소화하기 위해 매개변수를 조정하여 AI 모델이 패턴을 인식하고, 결정을 내리고, 출력을 생성하도록 가르치는 과정이다.