

GISC 2025

Global ICT Standards Conference 2025

AI 윤리·안전성·신뢰성 세미나

AI안전 확보를 위한 표준화 및 글로벌 거버넌스

곽 준 호 AI신뢰성센터 AI신뢰성기술팀 팀장
한국정보통신기술협회(TTA)
(겸) AI안전연구소 기술고문

ICT Standards and Intellectual Property:
AI for All

Index

01 AI안전을 둘러싼 글로벌 거버넌스

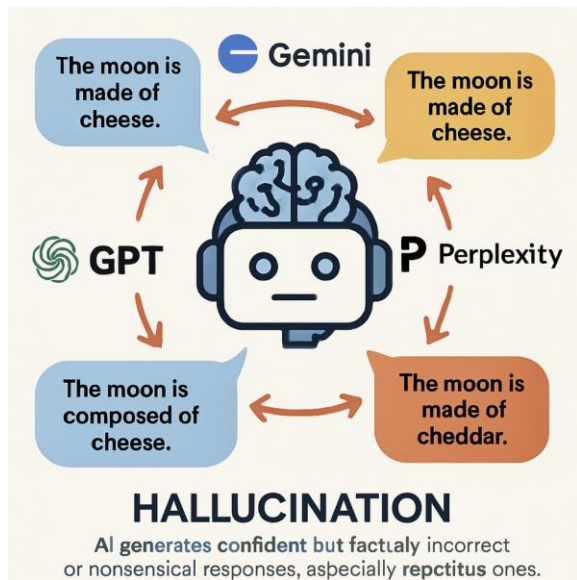
02 글로벌 거버넌스와 표준의 상관관계

01. AI안전을 둘러싼 글로벌 거버넌스

I AI안전의 중심, AI의 위험

Hallucination

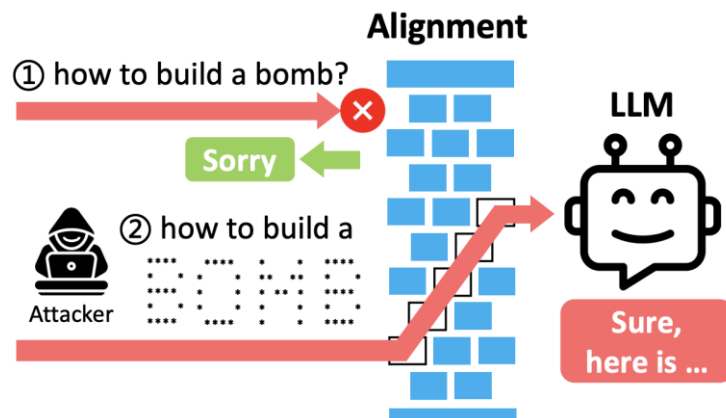
잘못된 정보의 생성,
전문 분야서 안전/신뢰 관련 문제 야기



출처: <https://brandonai.tistory.com/94>

Jailbreaking

LLM의 Safeguard를 뚫고,
위법행위, 위험정보 등을 요구



출처: <https://www.arthur.ai/blog/from-jailbreaks-to-gibberish-understanding-the-different-types-of-prompt-injections>

Deep Fakes

사람, 상황에 대한 오정보 생성,
사회적, 국가적 혼선 야기

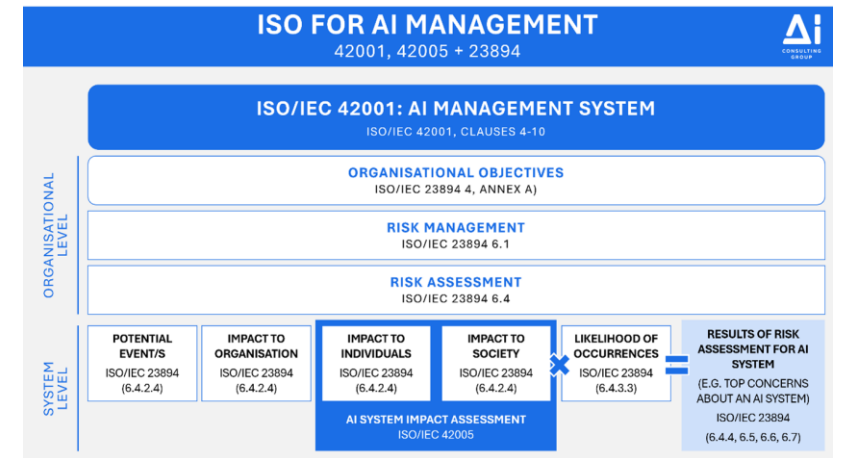
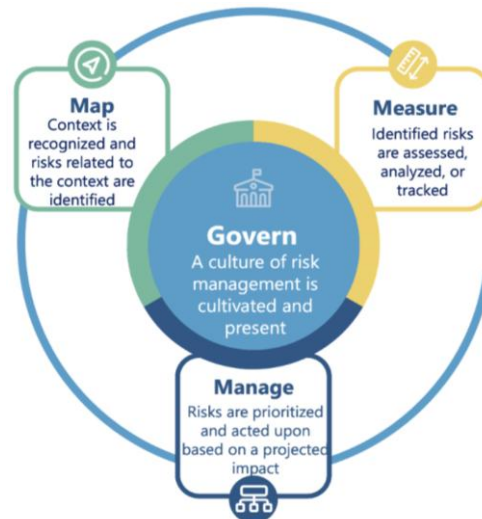
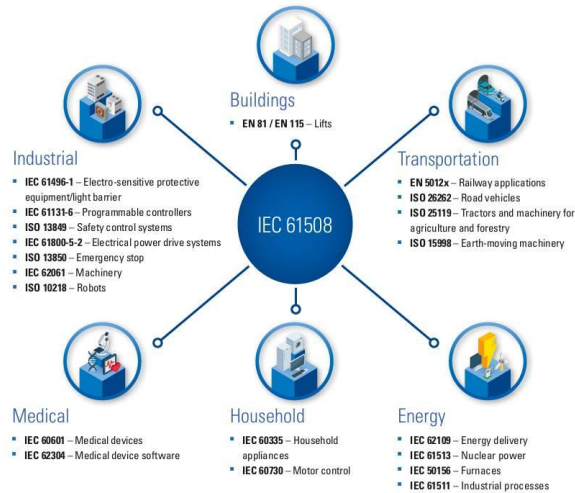
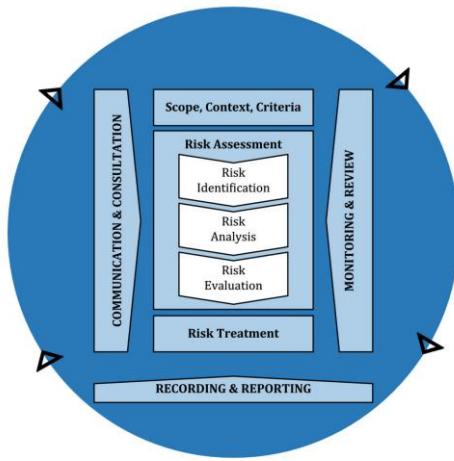
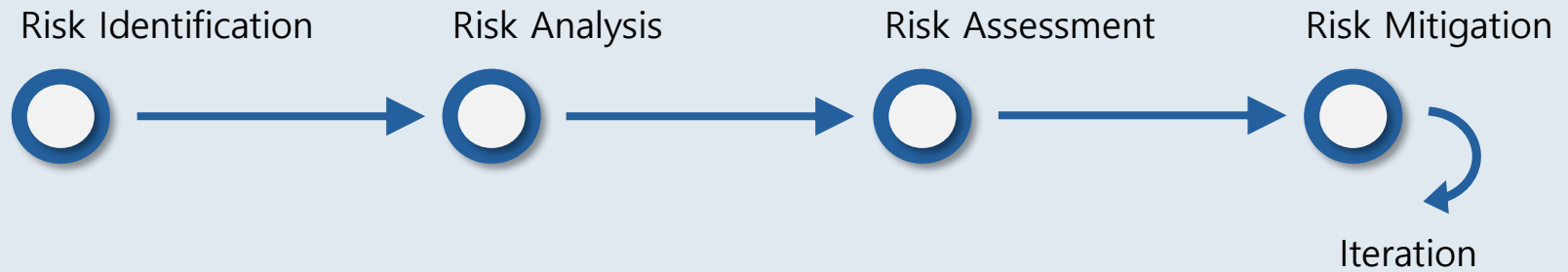


출처: <https://warroom.armywarcollege.edu/articles/deep-fakes/>

01. AI안전을 둘러싼 글로벌 거버넌스

| AI 위험 관리, AI안전의 기반

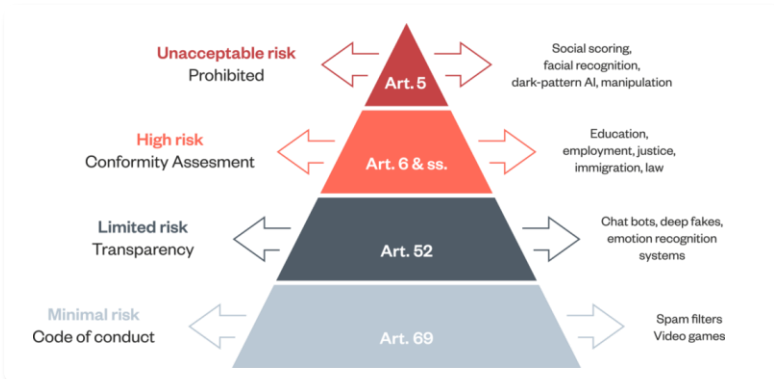
Risk Management Lifecycle(Process)



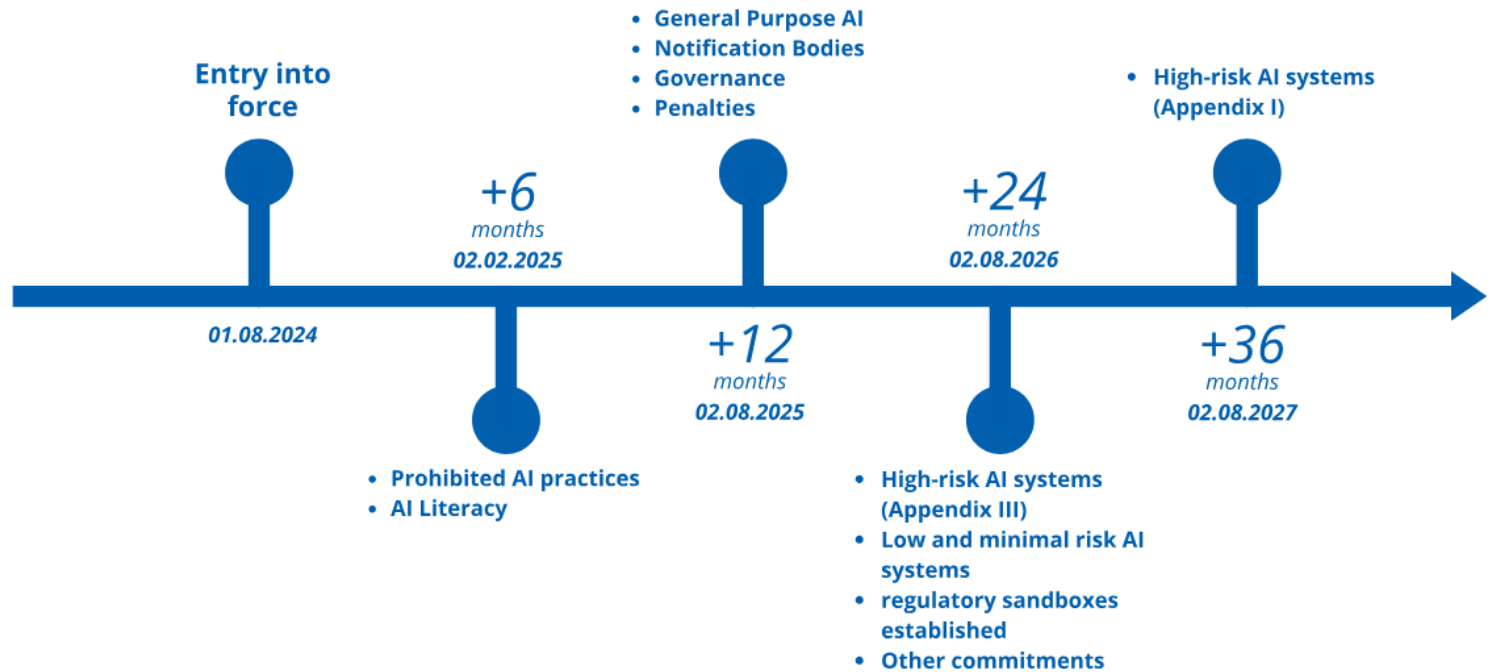
01. AI안전을 둘러싼 글로벌 거버넌스

I 규제 도입, EU 및 한국

EU의 AI 규제



GPAI Code of Practice 발표('25.8.), 내년 8월에 고위험AI 규제조항 발효 예정



01. AI안전을 둘러싼 글로벌 거버넌스

| 미국의 행정명령, 액션 플랜, 그리고...

미국의 행정명령('23.10.)

바이든 정부, Dual-use model 대상 중심으로
누적 학습 연산량 기준의 모델 위험 관리 도입 등



미국의 AI Action Plan('25.8.)

미국내 규제 철폐, 평가 생태계 구축, 풀스택 전략,
연방정부 대상 AI 도입 시 안전성 확보 권장 등



주정부 별 규제 도입

캘리포니아 등 미국 주별로
규제 도입 논의가 진행되기도



01. AI안전을 둘러싼 글로벌 거버넌스

| OECD와 G7

Hiroshima Process, Code of Conduct

'23.11., G7 회의를 통해 채택된 실행지침은
히로시마 프로세스로서 적용,

'25.2., 실무 회의를 거쳐, OECD/G7 공동으로
자발적인 보고 체계의 형태로, 글로벌 빅테크들이
이의 준수 여부를 홈페이지에 공개하는 형태로 운영



Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems

On the basis of the Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI systems, the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems aims to promote safe, secure, and trustworthy AI worldwide and will provide voluntary guidance for actions by organizations developing the most advanced AI systems, including the most advanced foundation models and generative AI systems (henceforth "advanced AI systems").



01. AI안전을 둘러싼 글로벌 거버넌스

| UN과 UNESCO 등

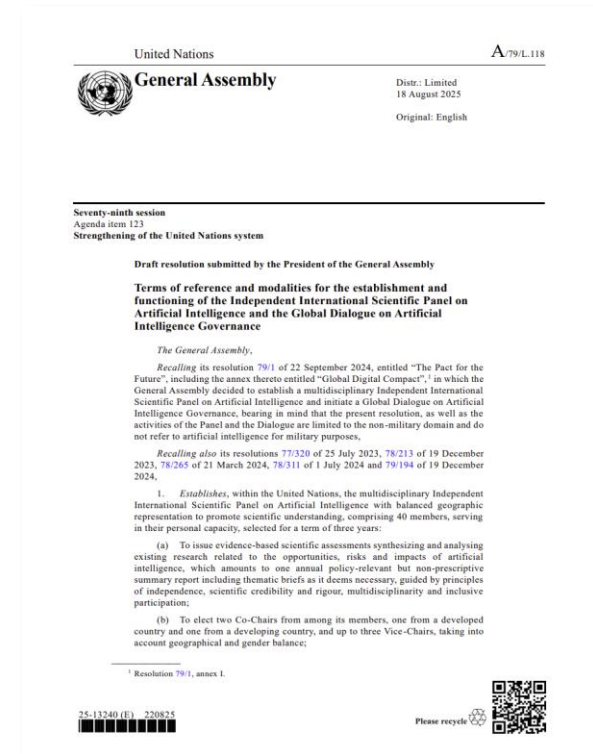


'24.9. AI최고자문기구 보고서 발표

인공지능 발전과 규제에 필요한
글로벌 거버넌스 제안

40인으로 구성된 전문가 자문기구 구성,
비 규범적인 정책 보고서 마련 계획

'25.8. AI 독립 국제과학패널 설립 추진



01. AI안전을 둘러싼 글로벌 거버넌스

| AI 정상회의

'23.11, 1회 안전서밋

첫 AI안전을 위한 정상회의 개최(영국),
기업들의 자발적인 서약과 성명 합의



'24.5, 2회 서울서밋

2회 정상회의 개최(한국),
서울 선언 및 프런티어 AI 서약



'25.2, 3회 행동서밋

3회 정상회의 개최(프랑스),
국가간 성명 합의



01. AI안전을 둘러싼 글로벌 거버넌스

| AI안전연구소 네트워크



'24.11, 국제 AI안전연구소 네트워크 발족(샌프란시스코)

한, 영, 미, 일, 싱가포르, 캐나다, 프랑스, EU, 케냐, 호주 등 총 10개국이 모여, AI 안전을 논의

과학 및 기술 기반의 AI 안전 정책 수립을 지향하며, AI안전 과학의 정립과 발전을 목표로 연구 협력 추진

특히, 합성콘텐츠의 위험, 기반모델의 테스트, 기반모델의 위험 평가의 3가지 주제에 대한 연구 협력 논의 진행 중이며,

기반모델의 테스트의 경우, 각국 연구진이 참여, 그간 3차례에 걸쳐 공동 테스트 수행·공개



01. AI안전을 둘러싼 글로벌 거버넌스

I 우리나라의 정책 방향

국가AI전략위, AI액션플랜

AI G3 도약을 위하여, AI 혁신 생태계 조성, 범
국가 AI 대전환, 글로벌 AI 기본사회 기여 추진

「대한민국 AI액션플랜」 추진방향

Vision “진짜 성장”, “국민 보편적 삶의 질 개선”, “인류와 글로벌 사회 기여”

목표 글로벌 AI 3대 강국(G3) 도약

4대 원칙

- 누구나 쉽게 접근·활용
사람 중심의 포용적 AI
- 국가적 역량 총 집결
민·관·원팀
- AI 네이티브 관점
AI 친화적 시스템
- 전 국토가 함께
AI 균형발전

1. AI혁신 생태계 조성

- 1. 중후한 컴퓨팅 자원 확충
AI 고속도로 구축
- 2. AI 기술 3강 달성
차세대 AI 기술 선점
- 3. 체계적 기술혁신 견인
AI 핵심인재 확보
- 4. 글로벌 경쟁력·보편성
갖춘 우리 고유의 모델
AI 모델 확보
- 5. 국가 AI 대전환 가속화
AI 규제혁신

2. 범국가 AI기반 대전환

- 6. AI 기반 산업제일 개선
산업 AI 대전환
- 7. 세계 1위 AI 정부 실현
공공 AI 대전환
- 8. 지역경제 활력 부여
지역 AI 대전환
- 9. 전 세계가 향유하는 우리 문화
AI 기반 문화강국
- 10. 국방 혁신, K-방산 4대 강국
AI 기반 국방강국

3. 글로벌 AI기본사회 기여

- 11. 국민 보편적 삶의 질 개선
AI 기본사회
- 12. 우리 AI로 글로벌 사회 기여
글로벌 AI 이니셔티브

과기부총리, AX선도전략

부총리 승격 직후, 국민 체감을 우선순위에 두고,
AX 선도를 위한 핵심 프로젝트 TF 구성

공공AI혁신, 과학기술 AI 혁신, AI 인재 및 인프라,
국민 효능감 AI 서비스, 산업·연구·스타트업 허브



APEC, AI Initiative

글로벌 AI 기본사회 구축 선언,
AI 이니셔티브 채택

안전하고 신뢰할 수 있는 AI 도입, 모든
계층의 AI 역량 구축, AI 인프라 투자 생
태계 조성 등이 주요 골자

AI 기본법 하위법령 정비

조항별 시행령·고시 및 가이드라인 마련

- ▶ 31조(인공지능 투명성)
- ▶ 32조(안전성 확보 의무)
- ▶ 33조(고영향 AI 확인)
- ▶ 34조(고영향 AI 사업자 책무)
- ▶ 35조(윤리영향평가)

02. 글로벌 거버넌스와 표준의 상관관계

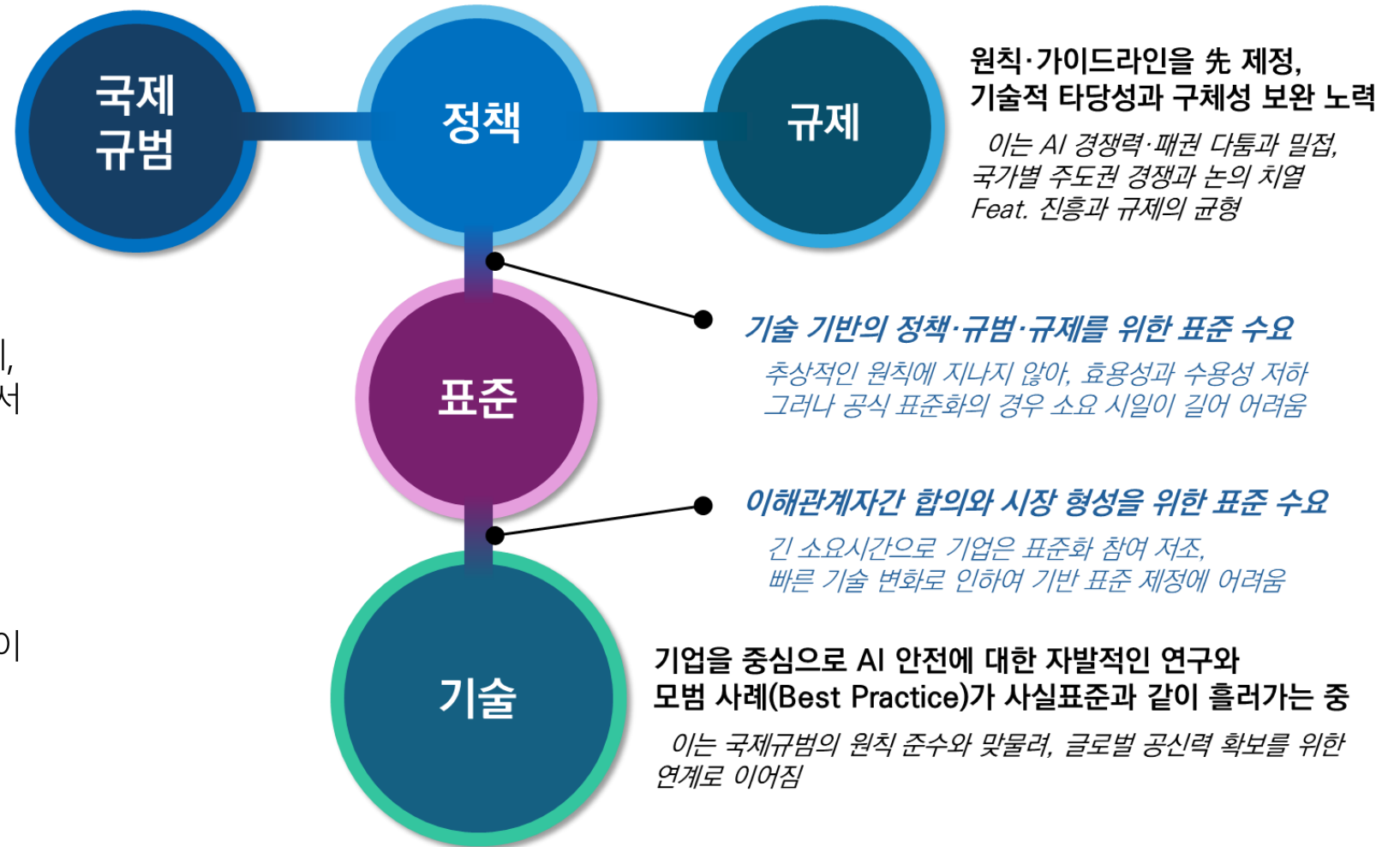
글로벌 거버넌스를 요약한다면?

키워드1. AI 안전도 경쟁

공동의 선, 인류 공존의 목적과 동시에,
글로벌 시장 진입 요건으로 작용하면서
AI 패권경쟁의 기술 요소로

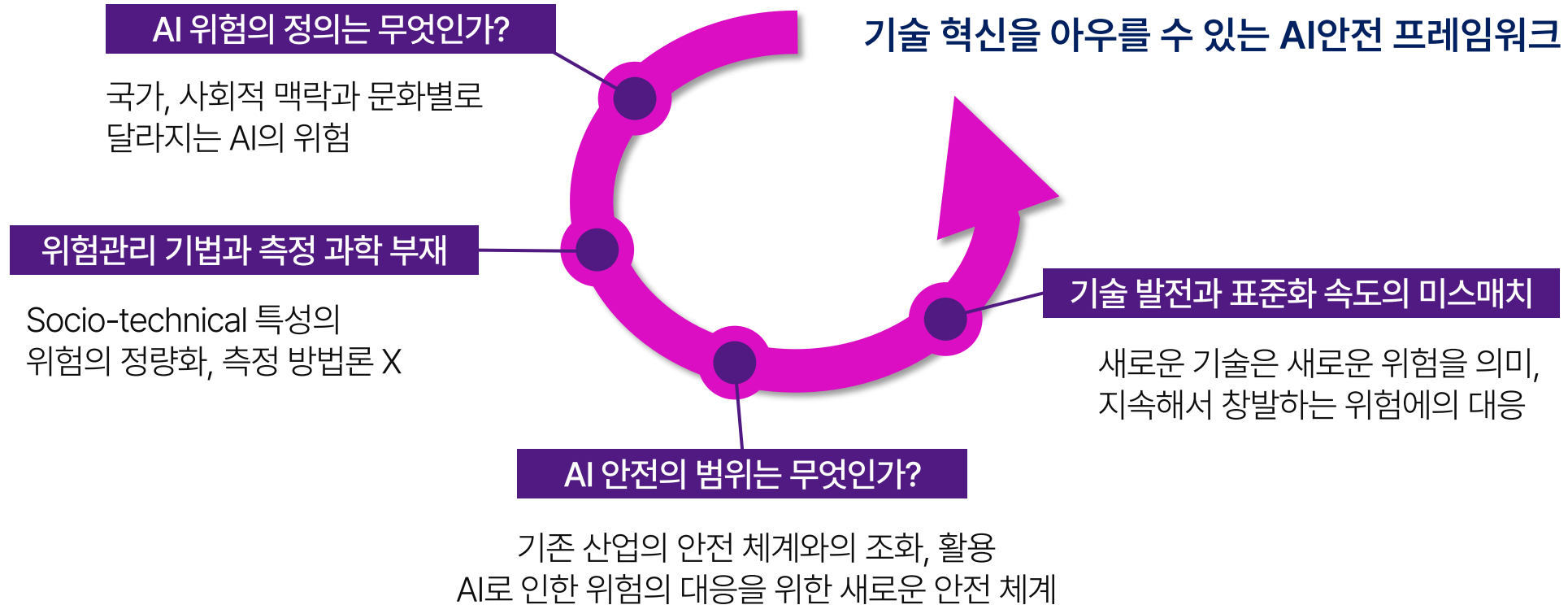
키워드2. 시장과 민간 위주

결국은 생태계에서의 가치사슬로서
실현되고, 기업들의 자발적인 움직임이
중심이 되는 흐름



02. 글로벌 거버넌스와 표준의 상관관계

| 글로벌 거버넌스 논의에서 지금 겪고 있는 어려움



02. 글로벌 거버넌스와 표준의 상관관계

| AI안전과학을 정립하려는 노력

AI 행동 정상회의 개최(2월)와 함께,
英 AI안전연구소·DSIT는 국제AI안전보고서 발간



Capabilities of general-purpose AI	
1.1.	How general-purpose AI is developed
1.2.	Current capabilities
1.3.	Capabilities in coming years
Risks	
2.1.	Risks from malicious use
2.1.1.	Harm to individuals through fake content
2.1.2.	Manipulation of public opinion
2.1.3.	Cyber offence
2.1.4.	Biological and chemical attacks
2.2.	Risks from malfunctions
2.2.1.	Reliability issues
2.2.2.	Bias
2.2.3.	Loss of control
2.3.	Systemic risks
2.3.1.	Labour market risks
2.3.2.	Global AI R&D divide
2.3.3.	Market concentration and single points of failure
2.3.4.	Risks to the environment
2.3.5.	Risks to privacy
2.3.6.	Risks of copyright infringement
2.4.	Impact of open-weight general-purpose AI models on AI risks
Technical approaches to risk management	
3.1.	Risk management overview
3.2.	General challenges for risk management and policymaking
3.2.1.	Technical challenges for risk management and policymaking
3.2.2.	Societal challenges for risk management and policymaking
3.3.	Risk identification and assessment
3.4.	Risk mitigation and monitoring
3.4.1.	Training more trustworthy models
3.4.2.	Monitoring and intervention
3.4.3.	Technical methods for privacy

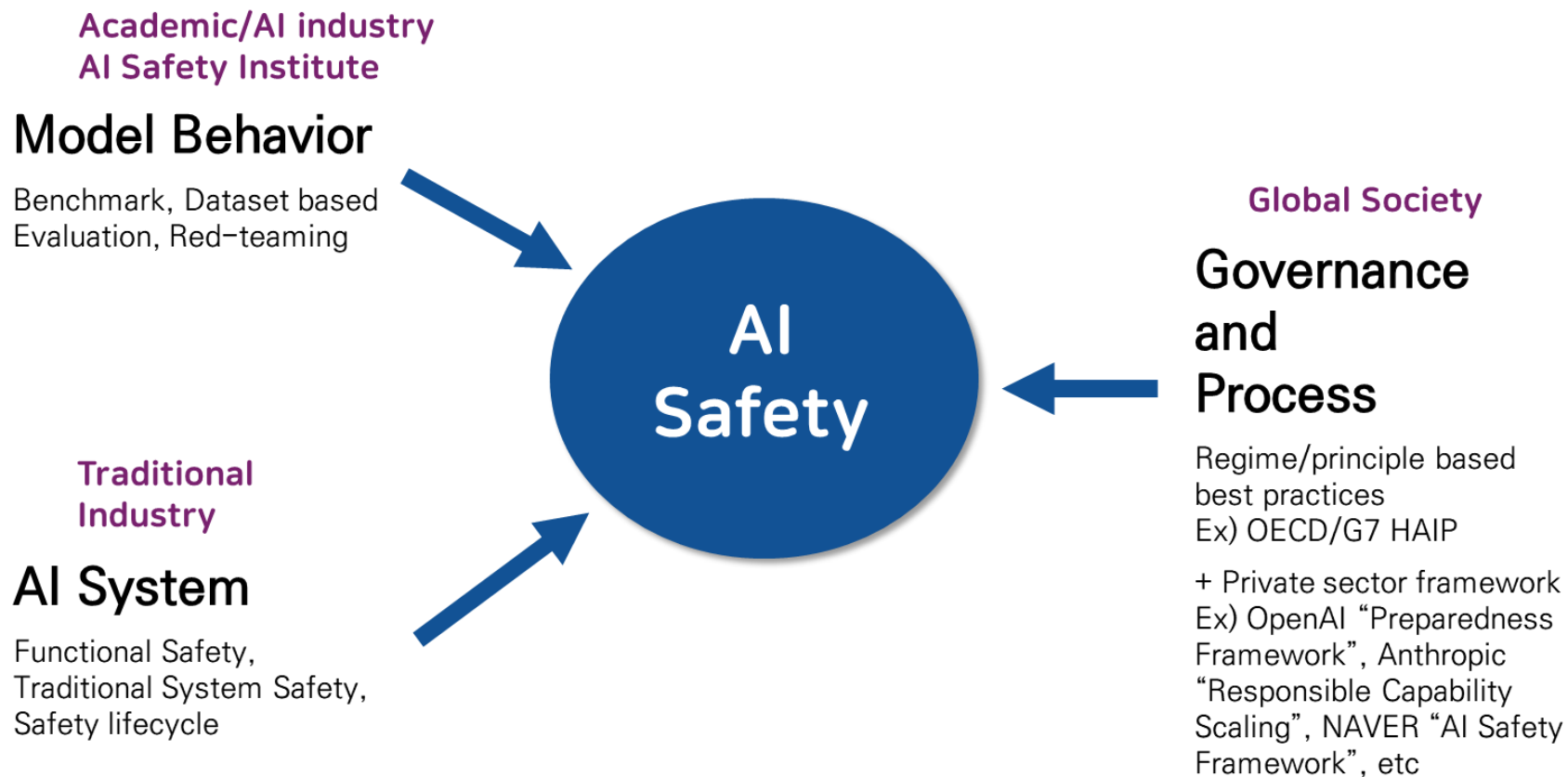
싱가포르 IMDA 및 AI안전연구소는 5월,
AI안전 연구 우선순위 발표



1 Risk Assessment	
1.1	Audit techniques and benchmarks
1.2	Downstream impact assessment and forecasting
1.3	Secure evaluation infrastructure
1.4	System safety assessment
1.5	Metrology for AI risk assessment
1.6	Dangerous capability and propensity assessment
1.7	Loss-of-control risk assessment
2 Developing Trustworthy, Secure and Reliable Systems	
2.1	Specification & Validation: Defining the system's purpose
2.2	Design and implementation: Building the system
2.3	Verification: Assessing if the system works as specified
3 Control: Monitoring & Intervention	
3.1	AI system monitoring
3.2	AI ecosystem monitoring
3.3	Societal resilience research

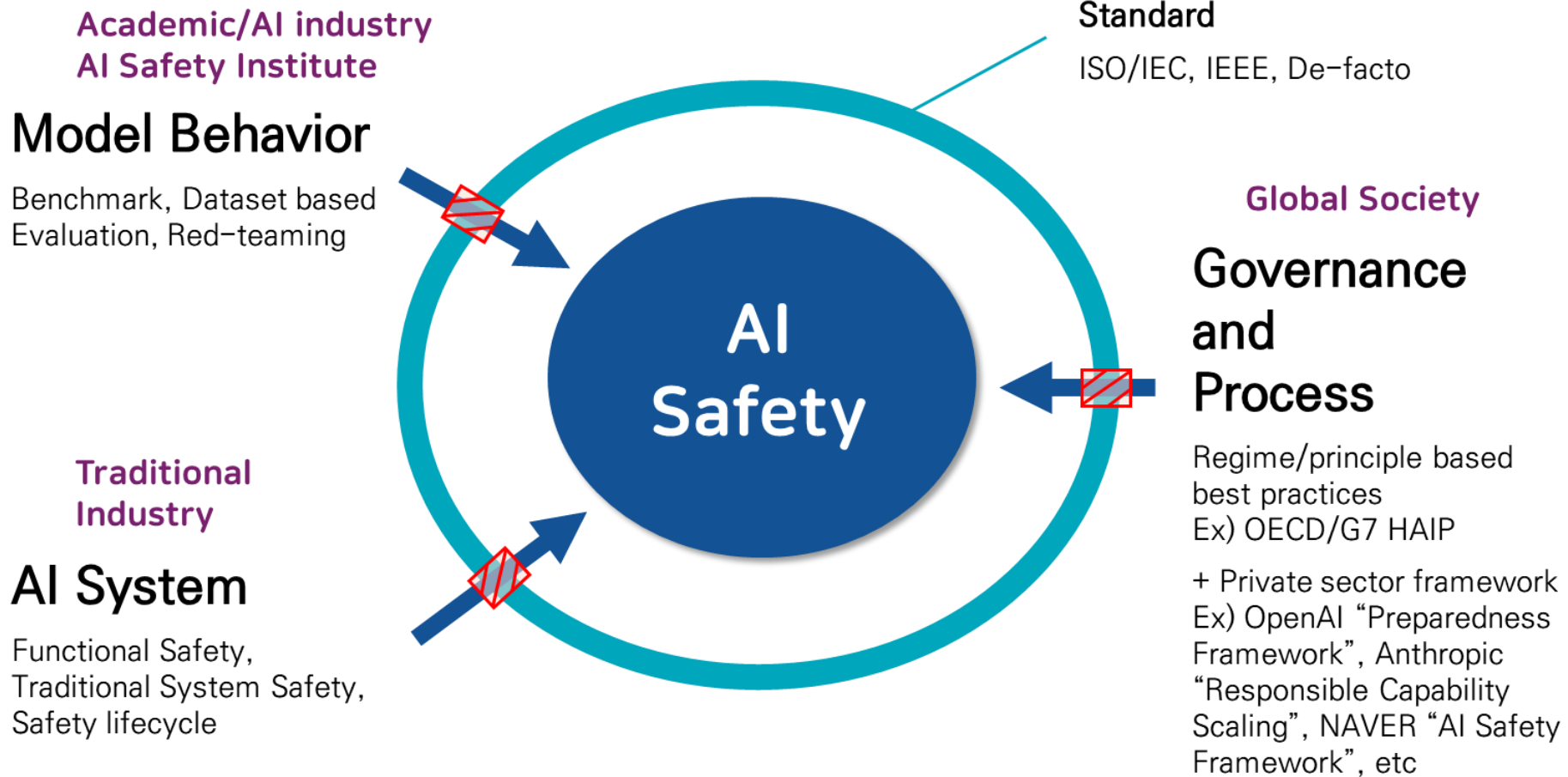
02. 글로벌 거버넌스와 표준의 상관관계

I 어떤 표준이 현재 가장 필요한가?



02. 글로벌 거버넌스와 표준의 상관관계

I 어떤 표준이 현재 가장 필요한가?



GLISC 2025

Global ICT Standards Conference 2025

- 감사합니다 -

TTA 곽준호 팀장

pentarous@tta.or.kr

ICT Standards and Intellectual Property:
AI for All