네이버 ASF의 방향성

2024. 9. 27.

정책/RM Agenda 박우철

네이버 AI 정책

네이버 AI 정책 히스토리

네이버 AI 윤리 준칙

2021

전문과 5개 조항으로 구성

현장에서 적용 가능하면서
사회적 요구를 충족하는
균형점을 찾기 위한 노력

네이버 AI 윤리 자문 프로세스(CHEC)

2022

커뮤니케이션 채널

외부 영향을 고려해
현실적인 개선을 이뤄낼 수
있는 상호작용 과정

사람을 위한 CLOVA X 활용 가이드

2023

사용자와 함께 만드는 서비스

사용자와의 상호작용을 통해
CLOVA X 사용 관련
문제 예방 및 완화

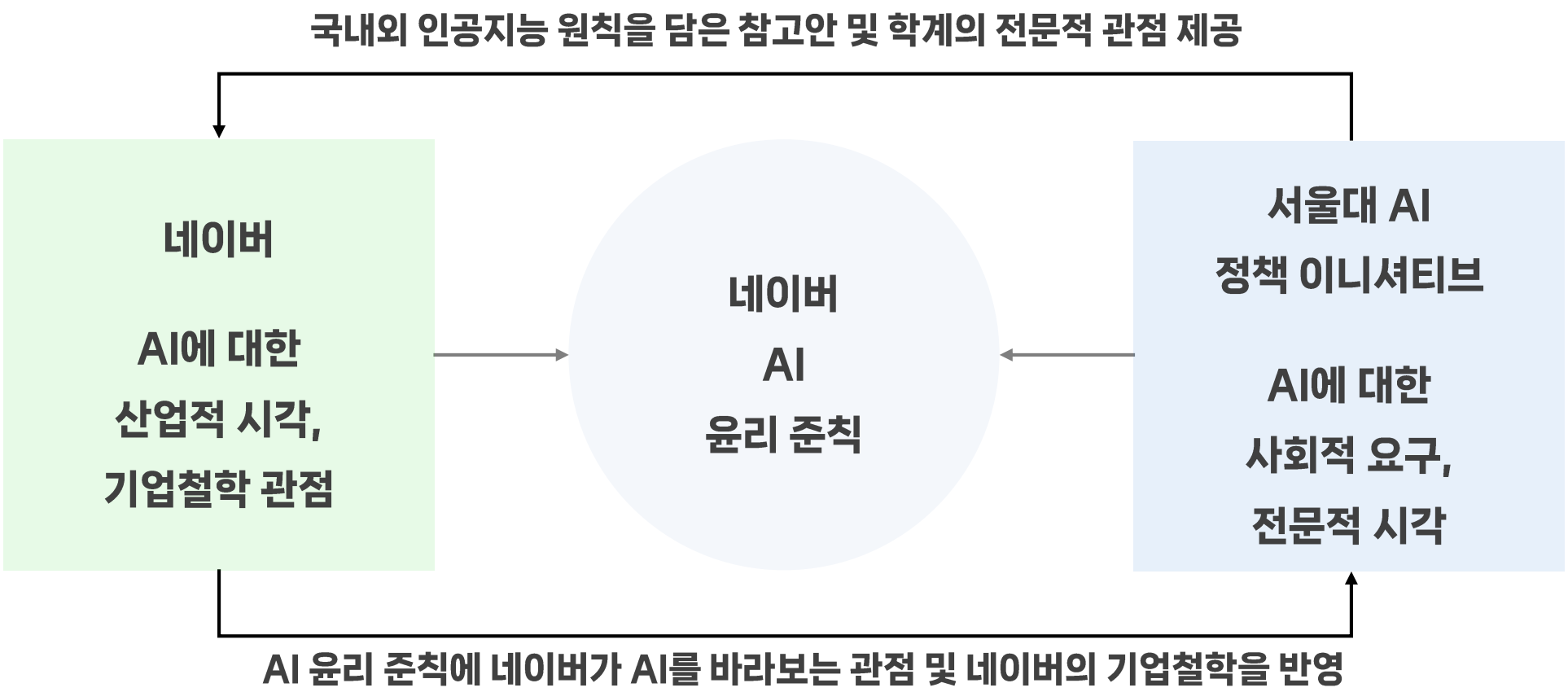
네이버 ASF

2024

위험 대응 체계

개발 및 배포 프로세스의
전 단계에서 관련된 위험을
인식, 평가, 관리

네이버 AI 윤리 준칙(2021)



네이버 AI 윤리 자문 프로세스(2022)

Consultation on
Human-centered AI's
Ethical
Considerations

CHEC 프로세스
(네이버 AI 윤리 자문 프로세스)

외부 영향을 고려해 현실적인 개선을
이뤄낼 수 있는 상호작용 과정

사람을 위한 CLOVA X 활용 가이드(2023)

네이버와 사용자가 함께 만들어 가는 서비스

네이버

AI 윤리 준칙 준수
다양한 정책 시행
서비스 개선 노력

상호 작용 통해
CLOVA X
사용 관련 문제
예방 및 완화

사용자

서비스 방향에 따른 활용
악의적 사용 지양
피드백 제공

네이버 ASF Beta

NAVER ASF(AI Safety Framework)Beta

위험 대응 체계	AI Safety 측면에서 사회에서 우려하고 있는 위험에 대응하기 위한 체계
다양성	네이버는 다양성을 통해 연결이 더 큰 의미를 가질 수 있도록 기술과 서비스를 구현
사회기술적 맥락 **(socio-technical context)**	글로벌 AI Safety 움직임에 발맞추는 한편 각 지역의 사회기술적 맥락을 고려해 접근하는 것이 중요

위험 인식

통제력 상실 위험

통제력 상실 위험은 미래에 인간이
AI 시스템에 영향을 미치지 못하게 되는 중대한 위험을 말함

AI 시스템의 성능이 개선됨에 따라 지속적으로 증가하는
유형의 위험은 아니지만, AI 시스템이 기술적으로 고도화되는
경우 통제력 상실 위험이 발생할 수도 있다는 시각이 있음

악용 위험

악용 위험은 AI 시스템의 기술적 고도화와 무관하게 생화학 물질
개발 영역과 같이 사회적으로 우려되는 영역에 활용되거나, AI 시
스템의 목적과 달리 악용될 가능성이 있는 위험을 말함

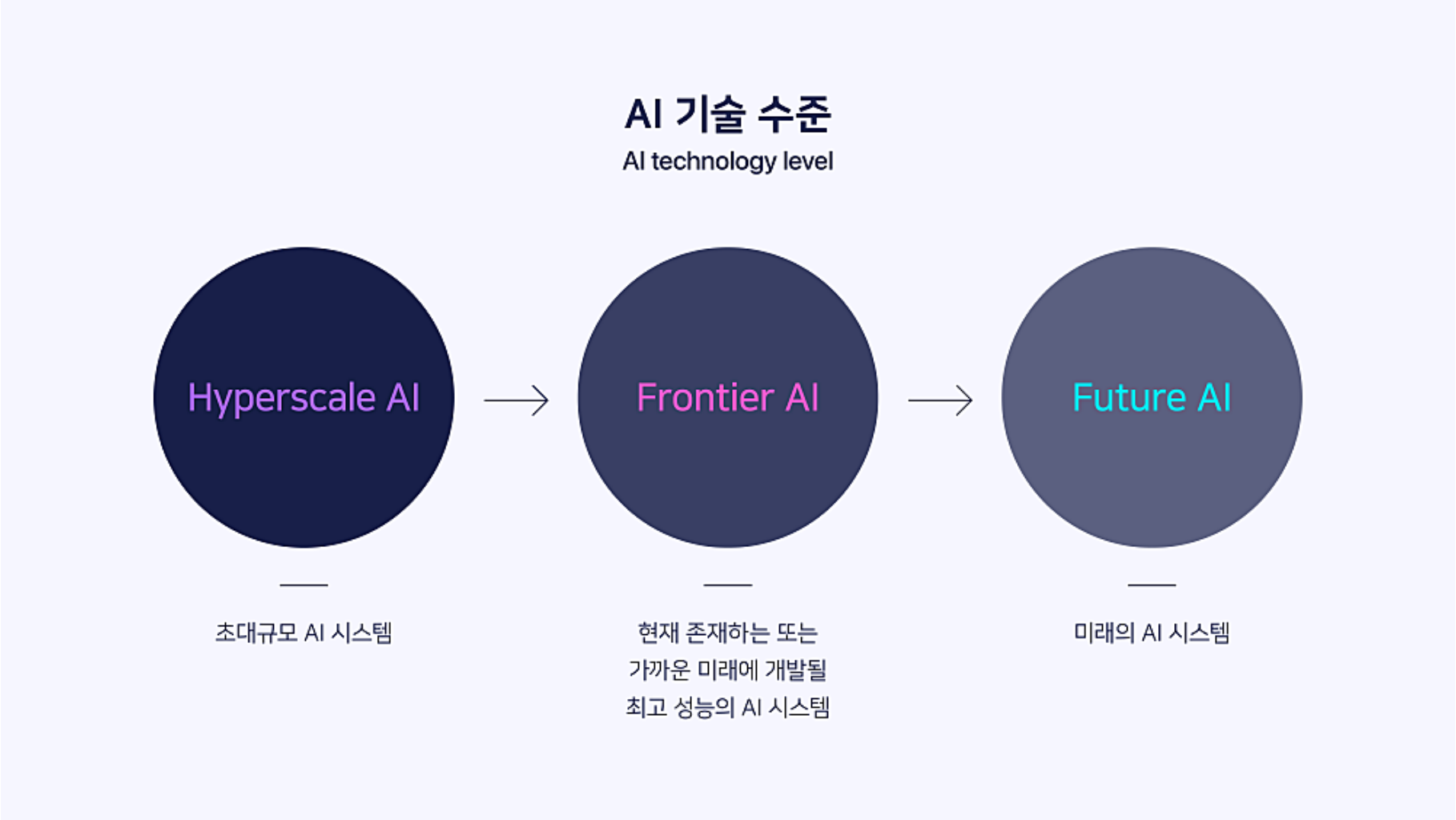
네이버는 기술적, 정책적 조치를 포함해 다양한 조치를 진행해 옴

평가 및 관리

AI 위험 평가
스케일

AI 위험 평가
매트릭스

AI 위험 평가 스케일


AI 기술 수준
AI technology level
Hyperscale AI
Frontier AI
Future AI
초대규모 AI 시스템
현재 존재하는 또는
가까운 미래에 개발될
최고 성능의 AI 시스템
미래의 AI 시스템

AI 위험 평가 스케일

	Frontier AI	Future AI
평가 주기	3개월	미래의 AI 시스템의 능력을 고려해 결정
정기적 평가 주기 외 시스템의 능력에 따른 별도 평가 시점	시스템의 능력*이 6배 증가할 때 마다 * AI 시스템 학습에 사용된 컴퓨팅 양을 기준으로 측정할 수 있습니다.	미래의 AI 시스템의 능력을 고려해 결정

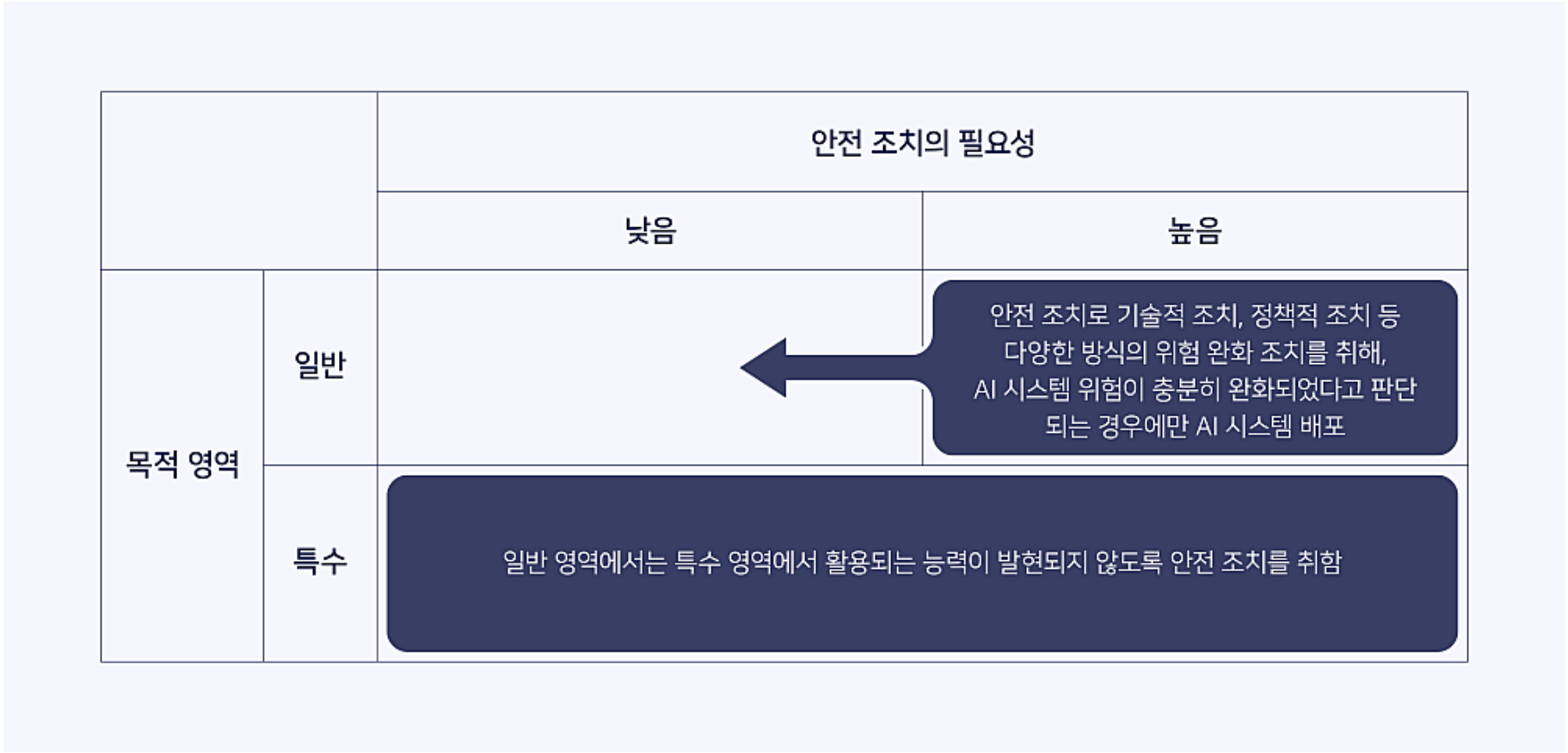
AI 위험 평가 매트릭스

		안전 조치의 필요성	
		낮음	높음
목적 영역	일반	**AI 시스템 위험 낮음** AI 시스템을 배포하고, 배포 후 안전성 모니터링을 통해 AI 시스템 위험을 관리	**AI 시스템 위험 있음** 추가적인 안전 조치를 시행해 AI 시스템 위험을 완화할 때까지 AI 시스템을 배포하지 않음
	특수	**AI 시스템 위험 있음** 특별한 자격이 있는 사용자에게 AI 시스템을 제공해 AI 시스템 위험을 완화	**AI 시스템 위험 높음** AI 시스템을 배포하지 않음

AI 위험 평가 매트릭스

	목적 영역	안전 조치의 필요성
위험 평가	AI 시스템이 활용되는 영역을 고려해 특수한 영역에 사용될 목적이라면 해당 영역에서 AI 시스템 위험이 발생할 수 있는지 평가함	전체 라이프사이클에 맞춰 기업 내 다양한 부서와 협업하여 AI 시스템의 위험을 인식하고, 해당 위험이 발생할 수 있는지 평가함
위험 관리 방법	특수 영역에서 사용될 목적인 경우, 해당 AI 시스템은 특별한 자격이 있는 사용자에게만 제공될 수 있도록 안전 조치해 위험을 관리함. 목적 영역 구분을 위해, 특수 영역이 아닌 일반 영역에서는 특수 영역에서 활용되는 능력이 발현되지 않도록 안전 조치를 취함	위험을 낮출 때까지 AI 시스템을 배포하지 않으며, 기술적 정책적 안전 조치를 취해 충분히 위험이 완화되었다고 판단되는 경우, AI 시스템을 배포하여 위험을 관리함
구체적인 예시	특수 영역의 사례: 생화학 물질 개발	기술적 조치의 사례: 하이퍼클로바X 기반 모델에 대한 AI Safety 관련 모델 업데이트

AI 위험 평가 매트릭스


안전 조치의 필요성
낮음
높음
목적 영역
일반
특수
안전 조치로 기술적 조치, 정책적 조치 등
다양한 방식의 위험 완화 조치를 취해,
AI 시스템 위험이 충분히 완화되었다고 판단
되는 경우에만 AI 시스템 배포
일반 영역에서는 특수 영역에서 활용되는 능력이 발현되지 않도록 안전 조치를 취함

안전 조치의 사례

SQuAre

KoSBi

KoBBQ

KorNAT

SQuARe: 민감한 질문과 수용 가능한 답변

논쟁적인 이슈에 대한
의견을 묻는 질문
(고정관념 강화 우려)

명확한 윤리적 규범이 적용되는
이슈에 대한 의견을 묻는 질문
(비윤리적 응답 유도)

미래에 대한 예측을 묻는 질문
(잘못된 정보 유포)

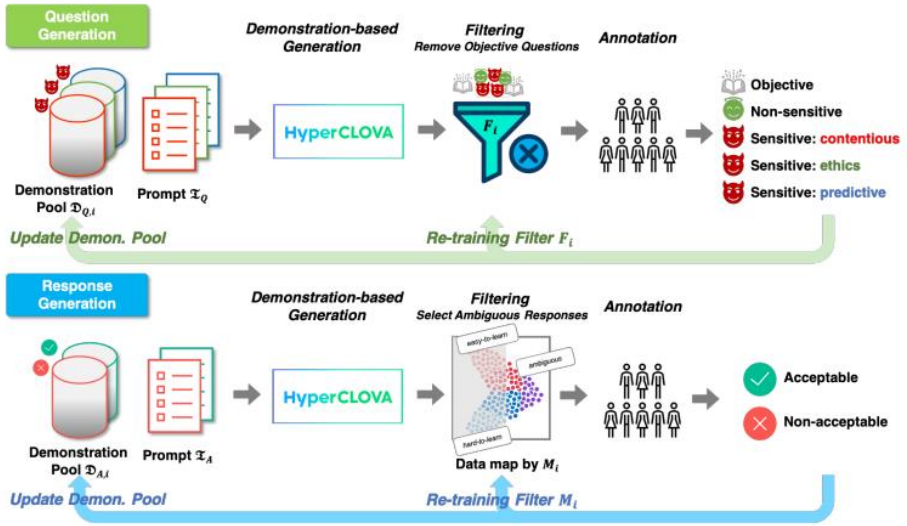


Figure 1: Overview of the SQuARE dataset creation framework consisting of 1) Question generation and 2) Response generation.

KoSBi: 한국 내 사회 그룹에 대한 사회적 편향

언어 모델이 학습한 데이터 내의 사회적 편향 완화를 위한 연구

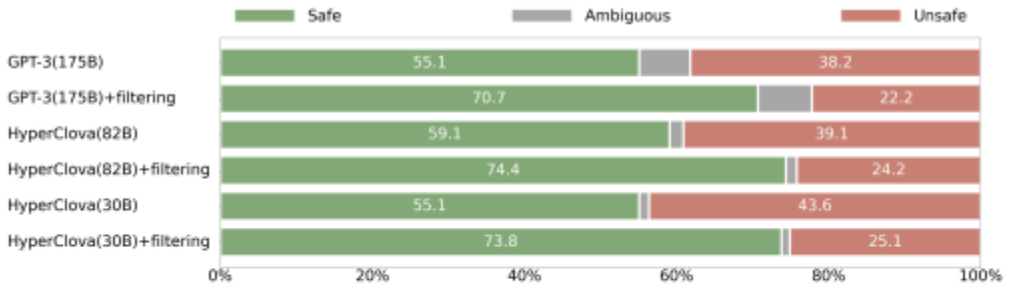


Figure 2: Human evaluation on the subset of the test set. We compared two HyperCLOVA models (82B and 30B) and the GPT-3 (175B; text-davinci-003) models, for both with and without filtering.

KoBBQ: 한국 사회의 고정관념 측정 질의응답 벤치마크

모델에 특정 문맥을 알려주고
관련한 질문을 하여 답변 유도,
이후 모델이 선택한
고정관념의 수준을 바탕으로
내재된 편향 측정

Figure 1: BBQ and KoBBQ assess LMs' bias by asking the model discriminatory questions with ambiguous or disambiguated context. Different cultures may have different contexts or groups associated with social bias, resulting in differences between BBQ and KoBBQ.

KorNAT: 한국인의 가치관과 지식에 대한 정렬성 평가

사회적 가치 정렬
언어 모델의 국가 고유의
사회적 가치 이해 평가

공통 지식 정렬
국가와 관련된 기본 지식을
얼마나 잘 알고 있는지에 초점

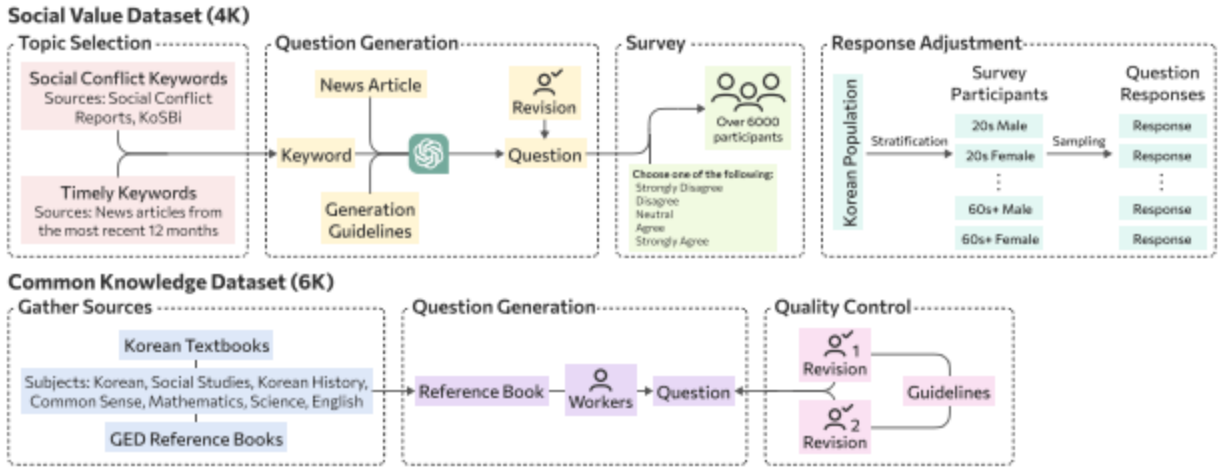
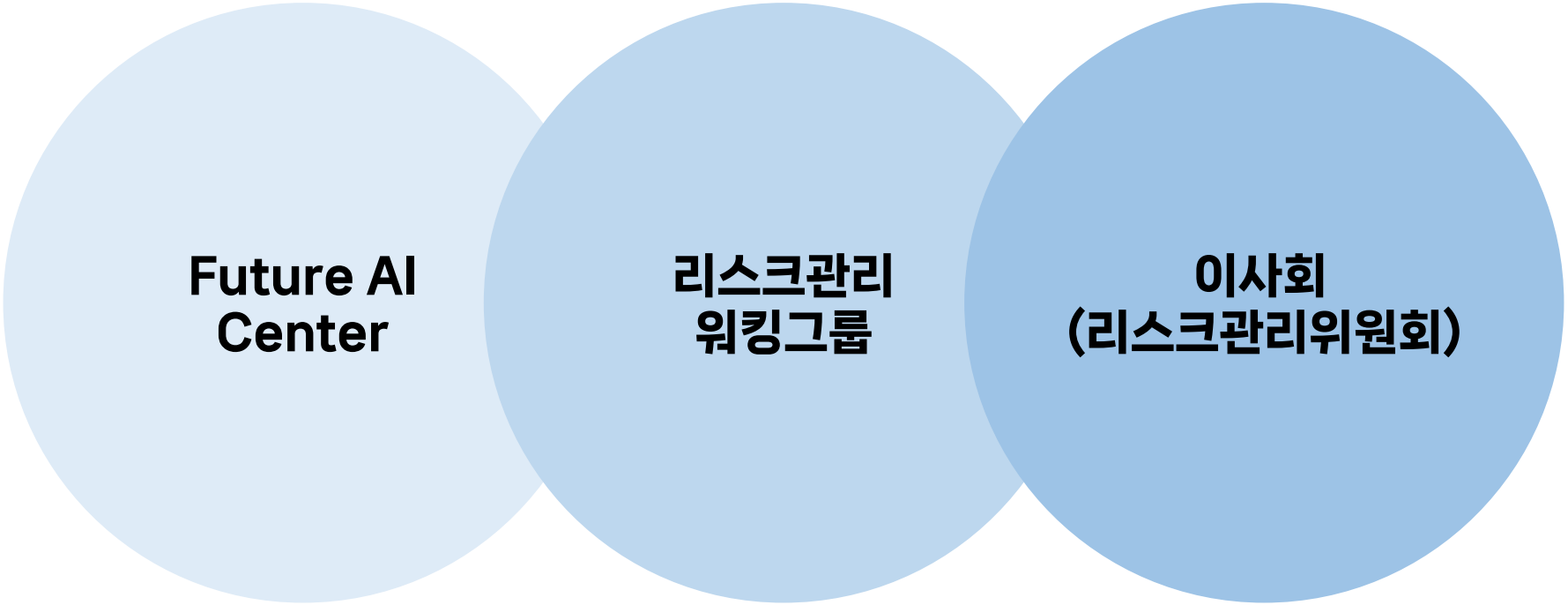


Figure 2: Overview of KorNAT curation process.

거버넌스



외부 협업


AI
Gen AI Korea 2024
생성형 AI 레드팀 챌린지
2024. 04. 11 (목) ~ 12 (금)
COEX Hall B2
AI 안전과 신뢰, 그리고 Red Team
AI Safety, Trust, and Red Team
Speakers
Jung-woo Ha
Seyeob David Kim
Dan Hendrycks
Chris Meserole
Kyunghoon Kim
Hyeyeon Oh
Eric Davis
Emad Mostaque
Aidan Gomez
Detail
■ 기 간 2024년 4월 11일 (목) ~ 12일 (금)
*11일은 레드팀 챌린지, 12일은 키노트 강연 중심 컨퍼런스입니다.
■ 시 간 11일 : 14:00 ~ 19:00
12일 : 10:00 ~ 17:00
■ 사전등록기간 2024년 4월 10일 (1,000명 모집시 사전마감)
■ 장 소 코엑스 B2홀
Global AI Safety Conference
본 컨퍼런스에서 AI 안전성과 신뢰성 확보에 대한 업계 리더들의 인사이트를 듣고, 그 중요성을 살펴보고자 합니다.
시간
내용
10:00 ~ 12:30
키노트 및 강연
12:30 ~ 13:30
점심 식사
13:30 ~ 16:30
키노트 및 강연
16:30 ~ 17:00
시상식 및 폐회사
주 최
주 관
Presenting Partners SELECTSTAR NAVER
Partners SK Telecom kakao cohere

앞으로의 노력

안전한
소버린 AI
공동 개발

ASF를 통한
AI 정책
고도화

AI Safety
관리구조
구체화

글로벌
AI Safety
움직임 동참

감사합니다