

AGI 안전 및 거버넌스 관행(practice)에 관한 국내외 전문가 인식 비교

정 선 화



본 보고서는 ETRI 기술정책연구본부 기본사업인
“국가 지능화 기술정책 및 표준화 연구 ” 를 통해 작성된 결과물입니다.

- 본 보고서의 내용은 연구자의 견해이며 ETRI의 공식 의견이 아님을 알려드립니다.

목 차

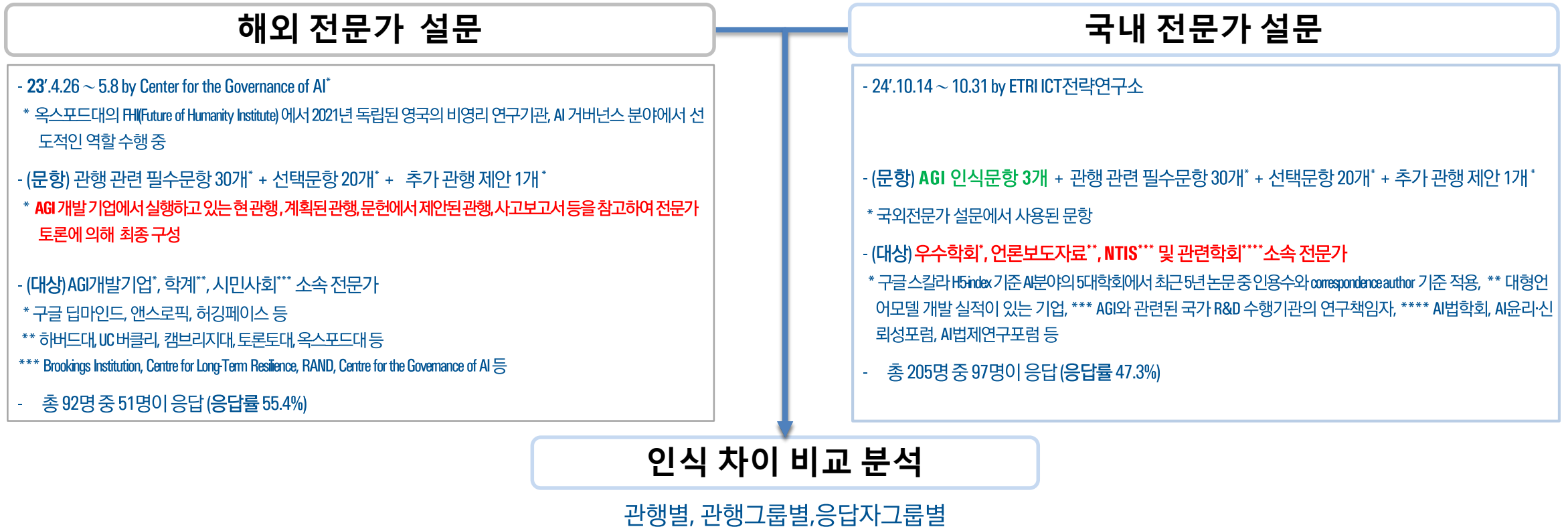
1. 연구 개요
2. AGI 안전에 대한 인식 수준 : 국내 전문가
3. 관행에 대한 국내외 전문가 인식 비교
4. 소결 : 국내 전문가 인식 조사, 국내외 전문가 인식 비교
5. 기대효과 및 활용방안

1. 연구 개요

◆ (목표) AGI 안전과 거버넌스 관행*에 관한 실무지침과 정책을 만드는데 실증적 근거 제공

* AGI 위험을 줄이기 위해서 **기업이 실행하는 Action Item**으로, 기업의 내부정책, 프로세스 및 조직구조와 관련된 사항으로 구성
(참고) 기업들은 다양한 **관행(Practice)**을 테스트 → 모범관행이 등장 → 정부 지침(Guideline) 및 표준에 반영 → 법(Act)으로 성문화

◆ (내용) AGI 생애 전주기에서 발생 가능한 위험을 줄이기 위한 기업의 관행에 관한 전문가 인식 조사 수행



[참고] 설문문항

AGI 안전인식 (3개 문항)

1. 인류존속을 위협하는 최대 위험
(5가지 예시 중 1개 선택)
2. AGI기술의 이점 vs 위험
(7점 리커트 척도)
3. AGI 기술의 위험 이유
(최대 3개 중복 선택)

(참고) 과기부 대국민설문, 영국 AISI의 설문, Rethink의 US 대국민 설문 참고

* 선택문항

**KYC(know-your-customer)

AGI 안전 및 거버넌스 관행에 관한 인식조사 (30개 필수, 20개 선택, 1개 주관식 문항)

개발 8개	Safety vs. capabilities	배포 8개	No unsafe open-sourcing	배포 4개	Monitor systems and their uses
	Pausing training of dangerous models		API access to powerful models		Report safety incidents
	Model containment		Treat updates similarly to new models		Emergency response plan
	Gradual scaling		KYC** screening*		Post-deployment evaluations*
	Pre-registration of large training runs		Treat internal deployments similarly to external deployments*	커뮤니케이션 10개	Publish alignment strategy*
	Dangerous capabilities evaluations		Avoid capabilities jumps*		Internal review before publication
	Tracking model weights*		Safety restrictions		Publish views about AGI risk
	Alignment techniques		Staged deployment		Publish results of internal risk assessments*
외부감사 7개	Third-party model audits	정보보안 6개	Security standards		Avoiding hype
	Red teaming		Military-grade information security		Publish results of external scrutiny*
	Increasing levels of external scrutiny		Security incident response plan*	Statement about governance structure*	
	Bug bounty programs		Protection against espionage*	Notify a state actor before deployment	
	Researcher model access		Industry sharing of security information*	Notify affected parties*	
	Third-party governance audits*		Dual control*	Notify other labs*	
	Inter-lab scrutiny*	위험관리 6개	Pre-deployment risk assessment	Enterprise risk management	
기타 1개	Pre-training risk assessment		Board risk committee*		
	Internal audit		Chief risk officer*		
			Background checks*		

2. AGI 안전에 대한 인식 수준 : 국내 전문가

- 인류의 존속을 가장 위협할 수 있는 위험 요소: 기후변화 > 핵전쟁 > AGI

핵전쟁	기후변화	AGI	소행성 충돌	전염병 대유행	합계
26	49	10	5	6	96
27.1%	51.0%	10.4%	5.2%	6.3%	100.0%

- AGI 기술의 잠재적 이점: 위험 < 이점

잠재적 위험이 많다	잠재적 위험이 다소 많다	잠재적 위험이 조금 많다	위험과 이점이 반반이다	잠재적 이점이 조금 많다	잠재적 이점이 다소 많다	잠재적 이점이 많다	합계
7	13	7	23	10	20	16	96
7.3%	13.5%	7.3%	24.0%	10.4%	20.8%	16.7%	100.0%



- AGI 기술의 잠재적 위험 (최대 3개 중복선택): 통제력 상실 > 허위정보 유포 및 여론 조작 > 이중용도 사용

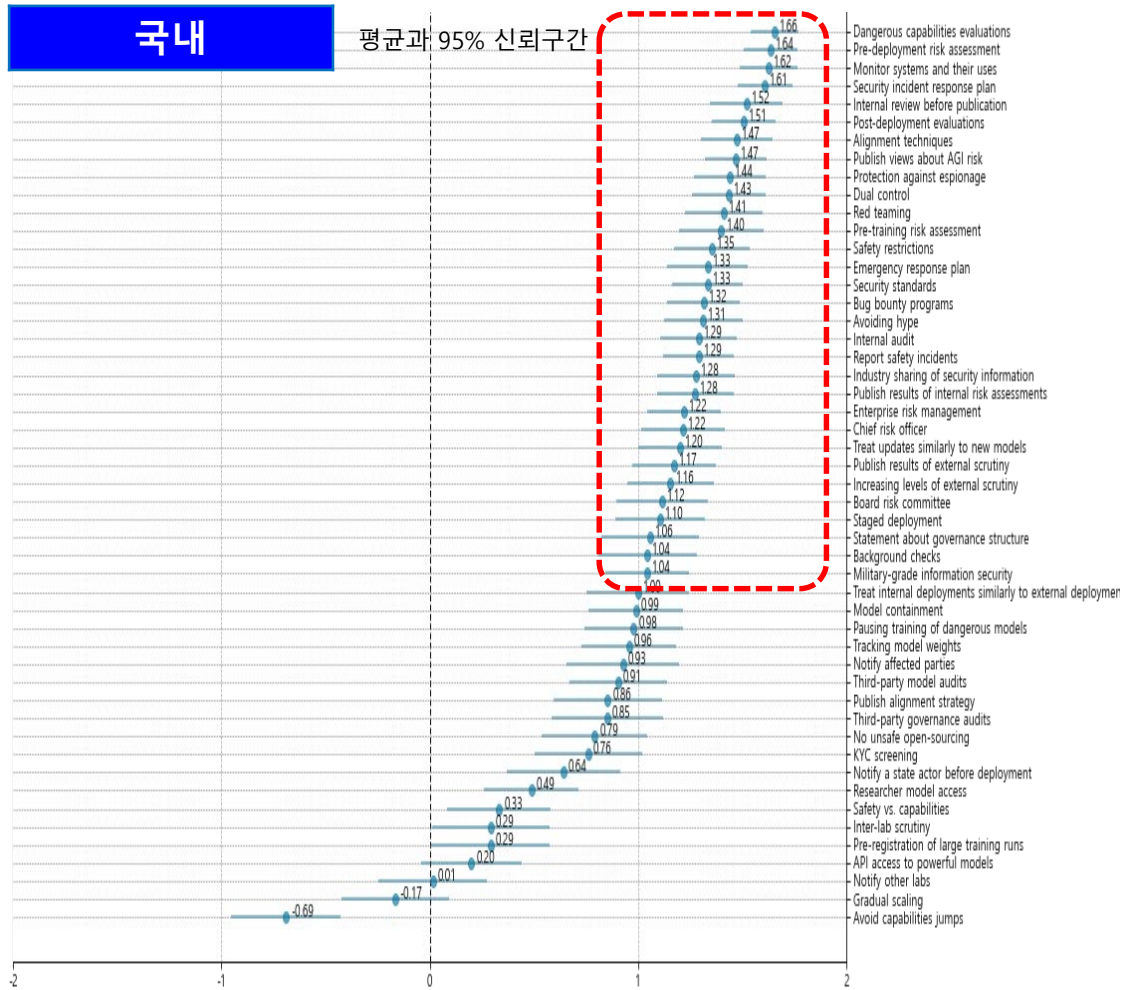
통제력상실	허위정보 유포 및 여론조작	이중용도 사용	가짜 콘텐츠로 인한 개인피해	사이버 공격	국가간 AI 기술 격차	편향성 및 대표성 부족	노동시장에 미치는 영향
20.7%	13.5%	11.7%	10.9%	10.9%	7.9%	6.0%	5.6%

- 이외, 시장독점 및 SPOF 위험(5.3%), 개인정보침해(3.8%), 제품기능 오류(1.5%), 환경위험(0.8%), 저작권침해(0.4%)

3. 관행에 대한 국내외 비교(1) : 동의 수준 - 평균 동의 기준

50개 항목에 대한 평균 동의 수준은 국내 1.05, 해외 1.39로 해외 전문가들의 관행 동의 수준이 매우 높았음

- 평균 동의 수준이 1 이상(약간 또는 매우 동의)인 관행 : 국내전문가는 32개, 해외전문가는 44개



(참고) 매우비동의 -2, 약간비동의 -1, 보통 0, 약간동의 1, 매우동의 2로 변환 후 문항응답의 평균 계산. 모르겠음 응답 제외

[참고] 관행별 전문가 평균동의 수준

관행그룹	관행(Practice)	해외전문가	국내전문가
개발	Alignment techniques	1.75	1.63
	Dangerous capabilities evaluations	1.88	1.66
	Gradual scaling	1.16	-0.17
	Model containment	1.33	0.99
	Pausing training of dangerous models	1.61	0.98
	Pre-registration of large training runs	1.06	0.29
	Safety vs. capabilities	1.69	0.33
	Tracking model weights	1.31	0.96
배포	API access to powerful models	1.19	0.20
	Avoid capabilities jumps	0.64	-0.69
	KYC screening	1.35	0.76
	No unsafe open-sourcing	1.25	0.79
	Safety restrictions	1.80	1.35
	Staged deployment	1.33	1.10
	Treat internal deployments similarly to external deployments	0.97	1.00
	Treat updates similarly to new models	1.14	1.20
배포 후	Emergency response plan	1.65	1.40
	Monitor systems and their uses	1.75	1.47
	Post-deployment evaluations	1.73	1.51
	Report safety incidents	1.72	1.29

(참고) 관행그룹내 관행은 알파벳순 정렬

(참고) 매우비동의 -2, 약간비동의 -1, 보통 0, 약간동의 1, 매우동의 2로 변환 후 문항응답의 평균 계산
모르겠음 응답 제외

관행그룹	관행(Practice)	해외전문가	국내전문가
외부 감사	Bug bounty programs	1.50	1.32
	Increasing levels of external scrutiny	1.58	1.16
	Inter-lab scrutiny	0.72	0.29
	Red teaming	1.76	1.41
	Researcher model access	1.22	0.49
	Third-party governance audits	1.34	0.85
	Third-party model audits	1.80	0.91
정보 보안	Dual control	1.43	1.43
	Industry sharing of security information	1.49	1.28
	Military-grade information security	1.40	1.04
	Protection against espionage	1.63	1.44
	Security incident response plan	1.74	1.61
	Security standards	1.48	1.33
	Security training	1.48	1.33
위험 관리	Board risk committee	1.39	1.12
	Chief risk officer	1.38	1.22
	Enterprise risk management	1.00	1.22
	Internal audit	1.29	1.29
	Pre-deployment risk assessment	1.90	1.64
	Pre-training risk assessment	1.65	1.33
	Risk management framework	1.65	1.33
커뮤니 케이션	Avoiding hype	1.18	1.31
	Internal review before publication	1.69	1.52
	Notify a state actor before deployment	0.90	0.64
	Notify affected parties	0.89	0.93
	Notify other labs	0.44	0.01
	Publish alignment strategy	1.50	0.86
	Publish results of external scrutiny	1.42	1.17
	Publish results of internal risk assessments	1.43	1.28
	Publish views about AGI risk	1.37	1.47
	Statement about governance structure	1.38	1.06
기타	Background checks	1.33	1.04

3. 관행에 대한 국내외 비교(2) : 동의 수준 상위 5개 - 평균 동의 기준

◆ 국내전문가는 활용될 때의 위험을 감소시키는 관행에, 해외전문가는 AGI의 철저한 위험평가에 높은 동의 수준을 보여줌

- 해외전문가의 경우 자체평가, 외부평가 등, 여러 측면에서 위험평가를 수행하여 최대한 안전한 AGI가 배포될 수 있도록 하는 관행에 동의가 높았으며, 동의 수준 또한 “매우 동의(2)”에 가까웠음
- 국내전문가의 경우 AGI가 활용될 때 발생할 수 있는 위험에 대응하는 관행에 동의 수준이 높았음

국내	그룹	동의 평균	설명
Dangerous capabilities evaluations	개발	1.66	자사 모델의 위해 가능성을 평가
Pre-deployment risk assessment	위험 관리	1.64	강력한 모델을 배포하기 전에 광범위한 위험식별, 분석 및 평가를 수행
Monitor systems and their uses	배포 후	1.63	배포된 시스템의 사용방식과 사회적 영향을 면밀히 모니터링
Security incidents response plan	정보 보안	1.61	보안사고에 대한 대응계획을 수립
Internal review before publication	커뮤니케이션	1.52	연구결과를 발표하기 전에 잠재적 위험을 평가하는 내부검토를 수행

해외	그룹	동의 평균	설명
Pre-deployment risk assessment	위험 관리	1.90	강력한 모델을 배포하기 전에 광범위한 위험식별, 분석 및 평가를 수행
Dangerous capabilities evaluations	개발	1.88	자사 모델의 위해 가능성을 평가
Third-party model audits	외부 감사	1.80	강력한 모델을 배포하기 전에 제3자 감사를 의뢰
Safety restrictions	배포	1.80	강력한 모델 배포 후 적절한 안전조치를 취
Red teaming	외부 감사	1.76	강력한 모델을 배포하기 전에 외부 레드 팀을 통한 취약점 분석을 수행

(참고) 관행 문장은 “AGI개발기관은 ~ 해야한다”로 표현됨

3. 관행에 대한 국내외 비교(3) : 동의 수준 하위 5개 - 평균 동의 기준

- ◆ 국내전문가는 개발과 배포에 관련된 관행에, 해외전문가는 커뮤니케이션에 관련된 관행에 낮은 수준의 동의를 보여줌
 - 해외전문가의 경우 하위 5개 동의에서도 동의 수준 중간이상 (>0)의 의견과, 구체적이지 않은 관행에 낮은 동의를 보임
 - 국내전문가의 경우 ‘Avoid capability jumps’와 ‘Gradual scaling’에 동의하지 않음을 표현하였으며, 기존 모델보다 능력이 현저히 뛰어난 모델과 대규모모델에 필요한 자원 문제에 허용적인 의견 표명

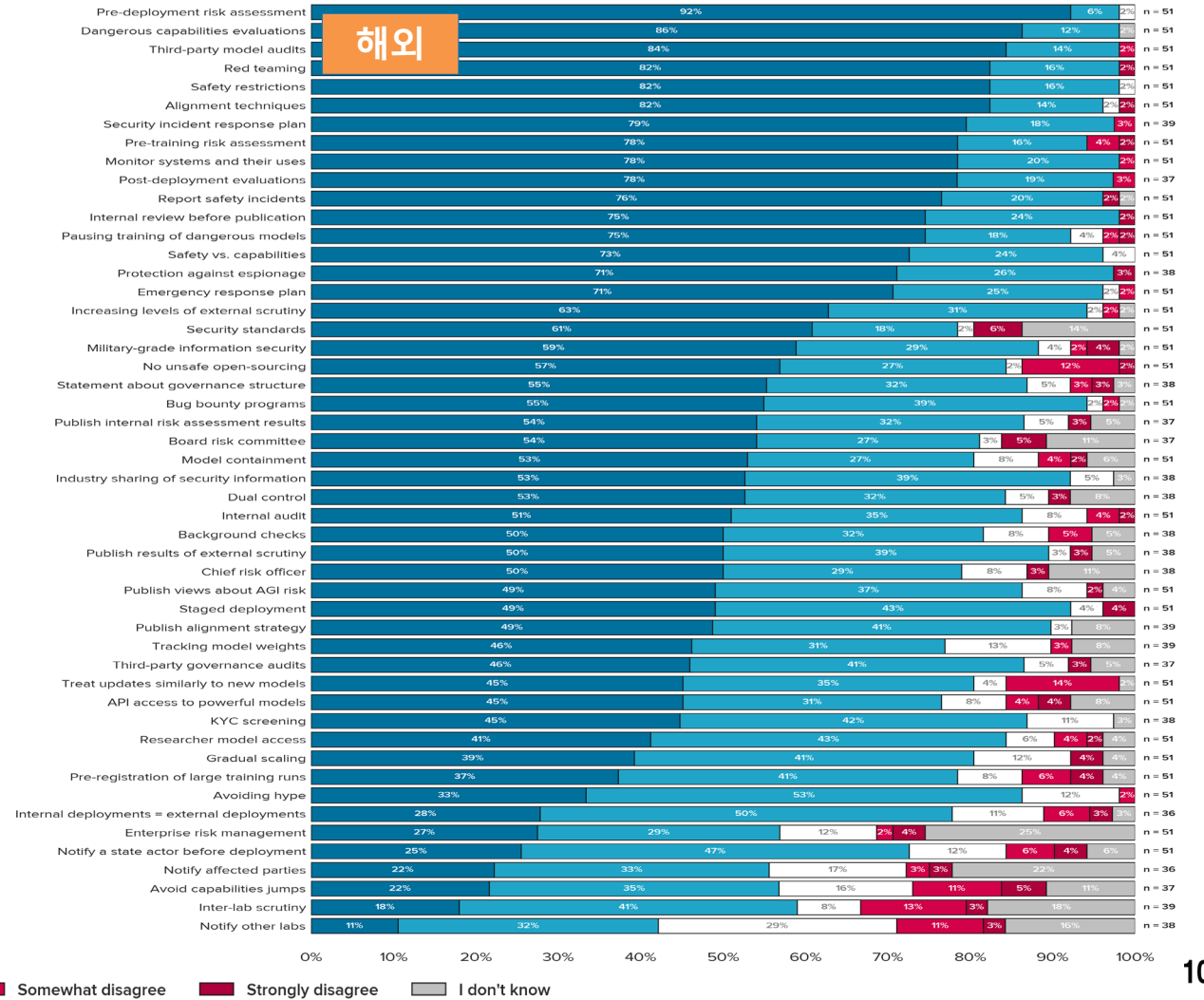
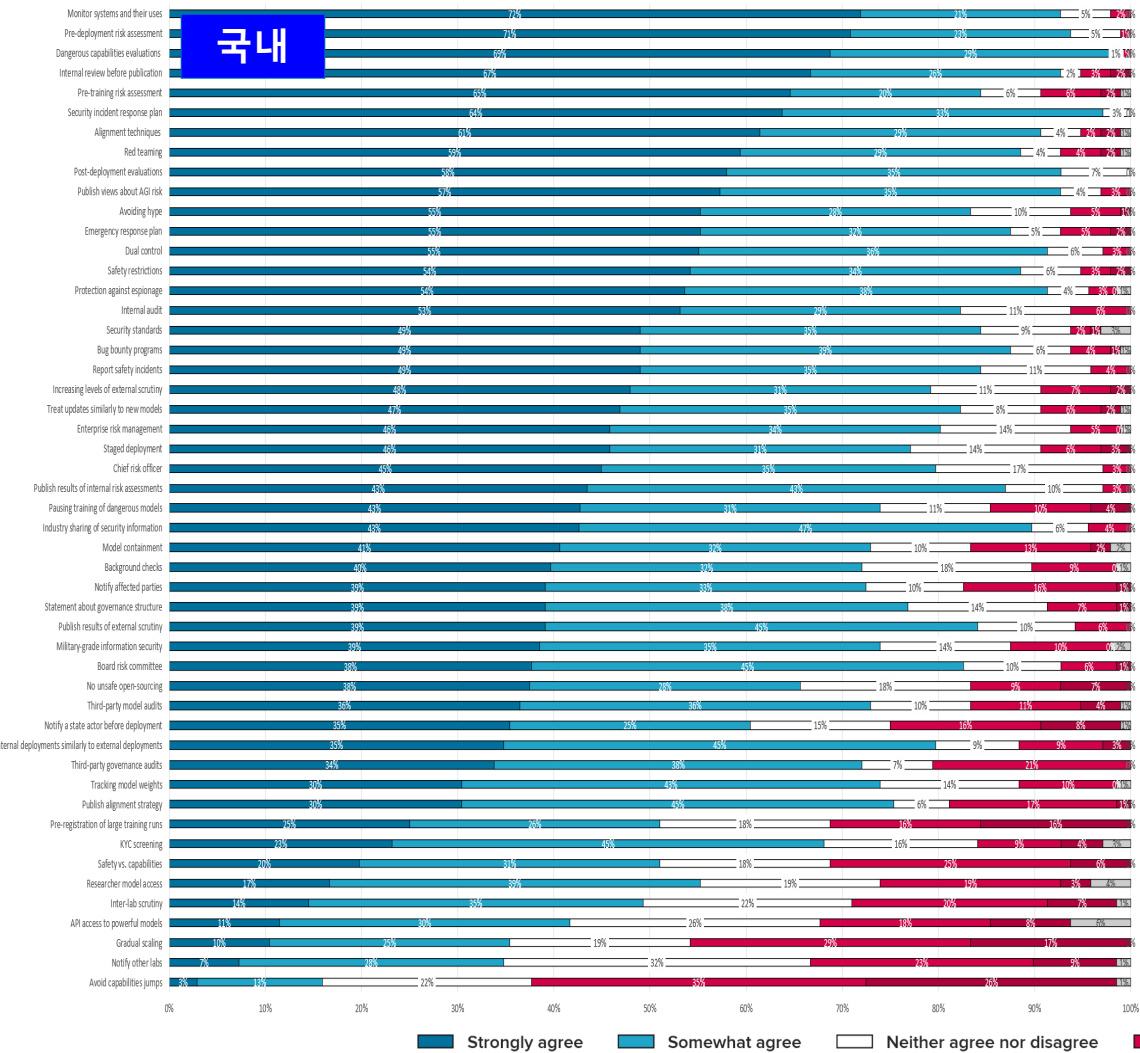
국내	그룹	동의 평균	설명	해외	그룹	동의 평균	설명
Avoid capabilities jumps	배포	-0.69	기존 모델들보다 능력이 현저히 뛰어난 모델을 배포 안	Notify other labs	커뮤니케이션	0.44	강력한 모델을 배포하기 전에 타 개발기관에 통보
Gradual scaling	개발	-0.17	최대 규모 모델의 훈련에 사용되는 컴퓨팅 자원과 데이터의 양을 점진적으로 증가	Avoid capabilities jumps	배포	0.64	내부 배포를 외부 배포와 유사하게 처리하고, 특히 배포 전 위험 평가를 수행
Notify other labs	커뮤니케이션	0.01	강력한 모델을 배포하기 전에 타 개발기관에 통보	Inter-lab scrutiny	외부 감사	0.72	타 개발기관의 연구자들이 배포 전에 강력한 모델을 검토할 수 있도록 허용
API access to powerful models	배포	0.20	강력한 모델을 오직 API를 통해서만 배포하는 것을 강력히 고려	Notify affected parties	커뮤니케이션	0.89	강력한 모델로 인해 부정적인 영향을 받을 수 있는 당사자들에게 통보
Pre-registration of large training runs	개발	0.29	일정 규모 이상의 모델 훈련을 관계 당국에 사전 등록	Notify state actor before deployment	커뮤니케이션	0.90	강력한 모델을 배포하기 전에 관계 당국 통보

(참고) 관행 문장은 “AGI 개발기관은 ~ 해야한다”로 표현됨

3. 관행에 대한 국내외 비교(4) : 동의 수준 - 동의 비율 기준

50개 관행 중 "약간 동의 또는 매우 동의"에 응답한 관행 : **국내 45개 관행**, **해외 49개 관행**

- 국내전문가는 동의를 유보하는(neither agree nor disagree) 비율이 높고, 해외 전문가는 모르겠음이 상대적으로 높았음



3. 관행에 대한 국내외 비교(5) : ‘중간 (neither agree nor agree)’ 응답

◆ 국내전문가의 경우 “중간”으로 답한 비율이 해외전문가보다 높아, 관행에 대한 이해도 제고 필요

- ‘Treat updates similarly to new models’과 ‘Chief risk officer’ 관행의 경우, 국내전문가는 ‘중간’과 함께 ‘모르겠음’ 답변 비율도 높아, 관행의 이해를 높이는 교육 및 홍보가 필요

국내	그룹	응답 비율	설명	해외	그룹	응답 비율	설명
Avoid capabilities jumps	배포	32%	기존 모델들보다 능력이 현저히 뛰어난 모델을 배포 안	Notify other labs	커뮤니케이션	29%	강력한 모델을 배포하기 전에 타 개발 기관에 통보
Treat updates similarly to new models	배포	26%	내부배포를 외부배포와 유사하게 처리해야하고, 특히 배포전 위험평가를 수행	Notify affected parties	커뮤니케이션	17%	강력한 모델로 인해 부정적인 영향을 받을 수 있는 당사자들에게 통보
Notify other labs	커뮤니케이션	22%	강력한 모델을 배포하기 전에 타 개발 기관에 통보	Avoid capabilities jumps	배포	16%	기존 모델들보다 능력이 현저히 뛰어난 모델을 배포 안
Gradual scaling	개발	22%	최대규모모델의 훈련에 사용되는 컴퓨팅 자원과 데이터의 양을 점진적으로 증가	Tracking model weights	개발	13%	강력한 모델의 가중치 복사본을 모두 추적
Chief risk officer	위험관리	19%	위험관리를 책임지는 고위임원인 CRO를 임명	Gradual scaling	개발	12%	최대규모모델의 훈련에 사용되는 컴퓨팅 자원과 데이터의 양을 점진적으로 증가

(참고) 관행 문장은 “AGI개발기관은 ~ 해야한다”로 표현됨

12%의 비율로 ‘중간’ 답변한 관행에는 Avoiding hype, Enterprise risk management, Notify state actor before deployment

3. 관행에 대한 국내외 비교(6) : ‘모르겠음(I don't know)’ 응답

🔥 해외전문가의 경우 “모르겠음”으로 답한 비율이 높은 관행들이 존재하는 한편, 국내전문가의 경우 상대적으로 매우 낮음

- 해외전문가의 경우 특히 구체적이지 않는 관행에서 ‘모르겠다’의 응답비율이 높았음
- 국내의 경우 익숙하지 않은 관행, 예를들어 CRO, 버그바운티프로그램, 과대광고 등에 ‘모르겠다’ 응답비율이 높았음

국내	그룹	응답 비율	설명	해외	그룹	응답 비율	설명
Treat updates similarly to new models	배포	6.3%	내부배포를 외부배포와 유사하게 처리해야하고, 특히 배포전 위험평가를 수행	Enterprise risk management	위험 관리	25%	기관전체의 위험관리프레임워크를 시행하고, 이는 AGI 맥락에 맞게 조정되고 개발기관이 사회에 미치는 영향에 집중
Chief risk officer	위험 관리	4.2%	위험관리를 책임지는 고위임원인 CRO를 임명	Notify affected parties	커뮤니케이션	22%	강력한 모델로 인해 부정적인 영향을 받을 수 있는 당사자들에게 통보
Bug bounty programs	외부 감사	3.1%	알려지지 않은 취약점과 위험한 능력을 보고한 자를 인정하고 보상하는 버그바운티 프로그램을 운영	Inter-lab scrutiny	외부 감사	18%	타 개발기관의 연구자들이 배포 전에 강력한 모델을 검토할 수 있도록 허용
Safety vs. capabilities	개발	2.9%	직원의상단 부분은 모델의 성능 향상보다는 안전성과 정렬개선에 집중	Notify other labs	커뮤니케이션	16%	강력한 모델을 배포하기 전에 타 개발 기관에 통보
Avoiding hype	커뮤니케이션	19%	AGI에 과도한 기대를 불러일으킬 수 있는 방식으로 강력한 모델을 공개하는 것을 피	Security Standards	정보 보안	14%	정보보안 표준을 준수해야하고 이 표준들은 AGI 맥락에 맞게 조정

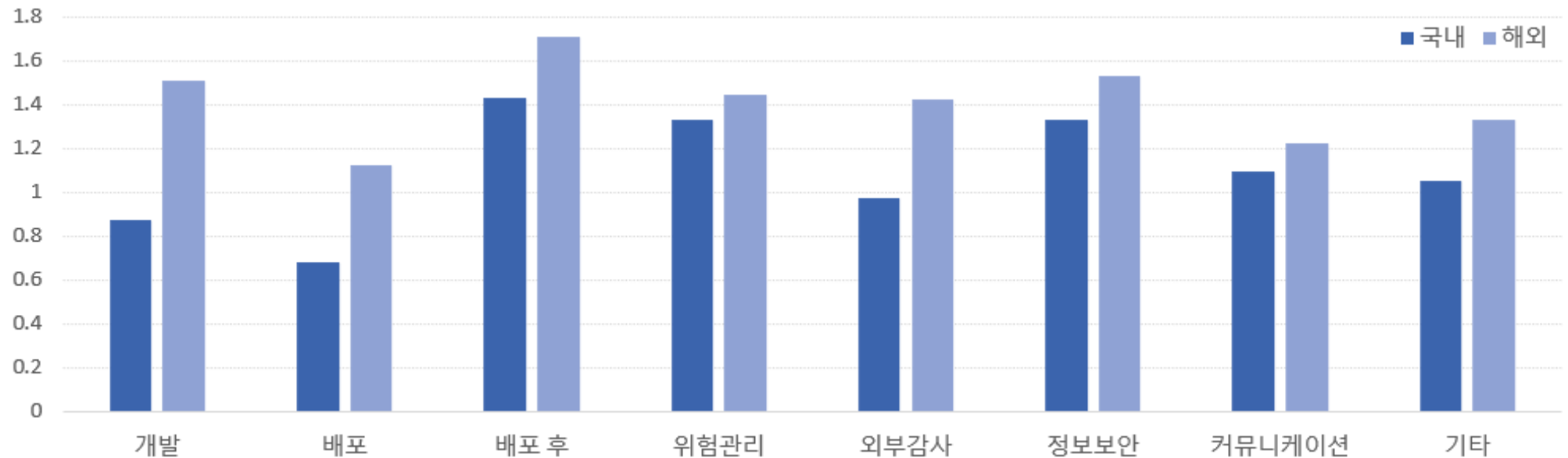
(참고) 관행 문장은 “AGI개발기관은 ~ 해야한다”로 표현됨

3. 관행에 대한 국내외 비교(7) : 관행그룹간 동의 수준

모든 관행그룹에서 해외전문가들의 동의 수준이 국내전문가들보다 높음

- 특히, 개발, 배포, 외부감사와 관련된 관행그룹에서 동의 수준의 차이가 크게 벌어짐

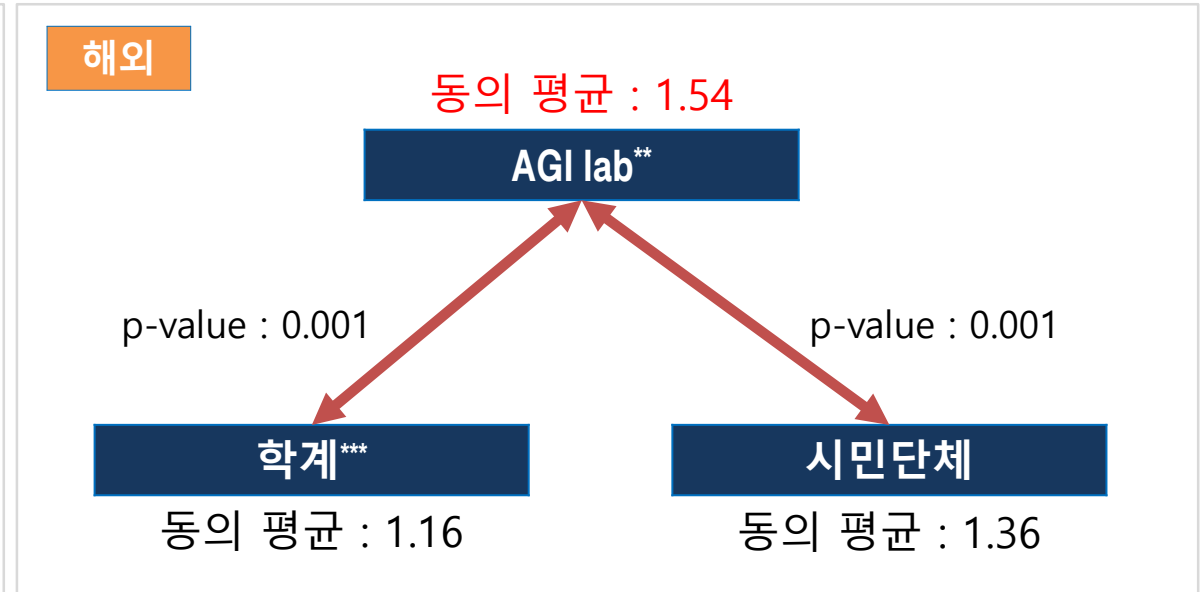
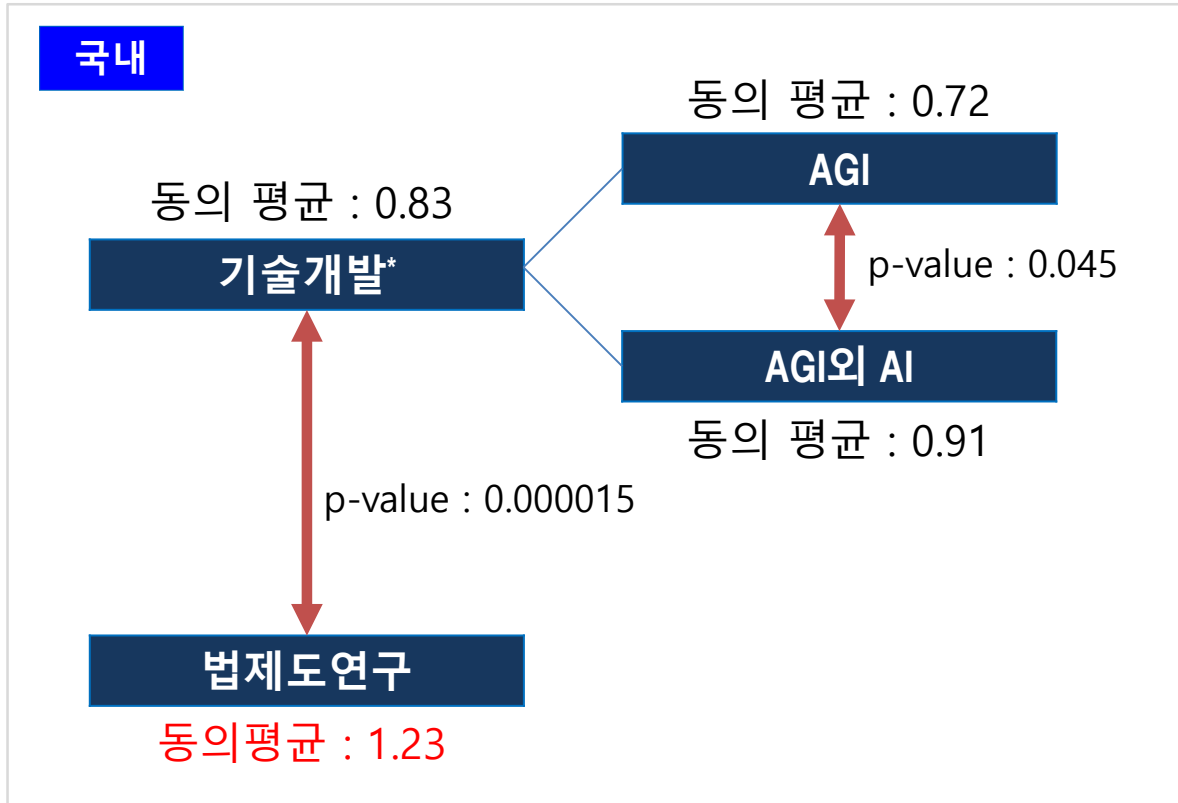
관행 그룹 구분	개발 (Development)	배포 (Deployment)	배포 후 (Post-Deployment)	위험관리 (Risk Management)	외부감사 (External Scrutiny)	정보보안 (Information Security)	커뮤니케이션 (Communications)	기타 (Others)
국내	0.87	0.68	1.43	1.33	0.97	1.33	1.09	1.05
해외	1.51	1.12	1.71	1.44	1.42	1.53	1.22	1.33
차이	-0.64	-0.44	-0.28	-0.11	-0.45	-0.20	-0.13	-0.28



3. 관행에 대한 국내외 비교(8) : 응답그룹간 관행 동의

◆ 국내는 기술개발과 법제도연구 그룹간 차이가 있었으며, 해외의 경우 AGI lab 그룹과 학계 및 시민단체와 각각 차이 존재

- 국내의 경우 기술개발 그룹의 동의 수준이 낮으나, 해외의 경우 AGI lab의 동의 수준이 상대적으로 높음



* AGI(또는 AI)를 개발하는 기업, 연구소, 대학 소속의 전문가로 구성

** AGI 개발을 목표로 하는 기업(OpenAI, 구글 딥마인드, 앤스로픽)과 LLM을 개발하는 기업(MS, 메타 등)

*** 옥스포드대 OpenMined, 토론토대 슈워츠-라이스만 기술사회연구소, 캠브리지대 미래지능센터, 캠브리지대 실존적위험센터 등에서 AGI(또는 AI) 안전 및 규제를 연구하는 전문가로 구성

◆ 해외 전문가의 경우 관행별로 응답그룹간 차이는 발생하지 않았음

4. 소결(1) : 국내 전문가 인식 조사

- ◆ AGI에 대한 국내 전문가의 인식 수준은 우호적인 편 → ◆ 안전 및 거버넌스 관행의 합의를 위한 노력이 더 필요
 - 인류존속을 위협하는 위험요인으로 기후변화(51.0%), 핵위기(27.1%) 보다 매우 낮은 10.4%가 동의
 - AGI 기술은 위협보다는 이점이 많다는데 48% 가 동의
- ◆ 법제도연구 그룹이 기술개발 그룹보다 관행에 더 동의 → ◆ 두 그룹 간의 인식 차이를 좁히기 위한 노력 필요
 - 기술개발 그룹의 동의 평균은 0.83, 법제도연구 그룹의 동의 평균은 1.23
 - 특히, 개발과 배포와 관련된 관행에서 두 그룹 간의 인식 차이가 존재하므로 차이의 원인을 규명하는 작업 필요
- ◆ 기술개발 그룹은 개발 및 배포와 관련된 관행에서 더 낮은 동의 → ◆ 우리의 기술적 상황과 맞추어 해외의 (표준이 되는) 관행과 차이 극복 필요
 - 개발과 배포에서 동의평균이 각각 0.87, 0.68로 다른 관행들보다 낮음
 - 선도기술을 추격하는 입장에서 AGI 기술혁신을 잠재적으로 저해할 수 있는 관행에 동의가 낮은 것으로 추정
- ◆ 동의 결정을 유보하는 응답 비율이 높음 → ◆ 원인규명과 함께 해당 관행의 이해를 높이는 노력 필요
 - 동의결정을 유보하는 원인 (정보 부족, 이해 부족 등)에 따라 교육, 홍보, 관행 표현의 명확화 등의 노력 강구
 - 배포와 외부감사와 관련된 관행에서 결정을 유보하는 비율이 높았음

4. 소결(2) : 국내외 전문가 인식 비교

- ▶ 해외전문가들의 관행 동의 수준이 국내전문가보다 높음 → ▶ 관행인식수준을글로벌수준으로 향상시키는 노력필요
 - 전체 관행에 대한 동의 평균에서 해외전문가는 1.39, 국내전문가는 1.05
 - 국내전문가는 32개의 관행에 대해서 평균 동의가 1 이상(약간 또는 매우 동의), 해외전문가의 경우 44개 관행에서 평균 동의가 1 이상
- ▶ 기술개발 그룹 동의 수준이 국내는 낮고 해외는 높음 → ▶ 국내 기술개발 그룹 낮은 동의 원인 파악 필요
 - 국내 기술개발 그룹 동의 수준 : 0.83 (< 전체 평균 1.05), 해외 기술개발 그룹 동의 수준 : 1.54 (> 전체 평균 1.39)
 - 국내의 경우 기술개발 그룹에 연구소, 대학의 전문가들이 포함된 반면, 해외의 경우 AGI(또는 AI)기술개발 기업이 응답하였음에도 불구하고 국내외 그룹 모두와 비교하여도 가장 높게 동의
- ▶ 관행별로 국내외 인식 차이 존재 → ▶ 관행의 표준화 및 정책 우선순위 결정 때 고려 필요
 - 국내전문가는 활용될 때의 위험을 감소시키는 관행에, 해외전문가는 AGI의 철저한 위험평가에 높은 동의 수준을 보여줌
 - 국내전문가는 개발과 배포에 관련된 관행에, 해외전문가는 커뮤니케이션에 관련된 관행에 낮은 수준의 동의를 보여줌
- ▶ 국내 전문가가 해외보다 중간으로 응답한 비율이 높음 → ▶ 관행의 이해도 제고를 위한 교육과 홍보 필요
 - 'Treat updates ~ models'와 'Chief risk of officer' 관행에서 국내의 "중간" 응답 비중이 특히 더 높음

5. 기대효과 및 활용방안

AGI 개발기업은 자체 실행하고 있는 관행과의 차이분석을 통해 관행을 개선하고 실행 관행 확대

- 50개의 관행 중 일부는 이미 AGI 개발기업에서 실행하고 있는 관행도 있지만, 대다수는 아직 미 실행
예) 'Dangerous capabilities evaluations' 관행은 OpenAI에서 GPT-4출시 때 실행
예) 'pre-training risk assessment' 관행을 실행하는 기업은 조사되지 않음
- 전문가들의 높은 동의는 AGI 개발기업들이 앞으로 추가적으로 관행을 실행할 수 있도록 독려

AGI 안전 관행의 표준화 과정과 정책 수립에 정보 제공

- AI 안전연구소에서 규제 의 우선순위를 결정하는데 참고자료로 활용
- 국내외 전문가 인식 차이를 고려하여 글로벌 관행을 논의할 때 국내 전문가의 의견을 파악하는 자료로 활용

AGI 관행은 점점 더 범용적인 AI 시스템을 개발하고 배포하는 AI 개발 기업에도 적용 가능

영문 축약	[한글 축약] 관행
Alignment techniques	[정렬 기술 적용] AGI 개발기관은 최신 안전 및 정렬 기술을 구현해야 한다.
API access to powerful models	[강력한 모델의 경우 API로만 접근] AGI 개발기관은 강력한 모델을 오직 API를 통해서만 배포하는 것을 강력히 고려해야 한다.
Avoid capabilities jumps*	[급격한 능력 도약 모델의 배포 기피] AGI 개발기관은 기존 모델들보다 능력이 현저히 뛰어난 모델을 배포해서는 안 된다.
Avoiding hype	[과장 광고 자제] AGI 개발기관은 AGI에 대한 과도한 기대를 불러일으킬 수 있는 방식으로 강력한 모델을 공개하는 것을 피해야 한다.
Background checks*	[이력 검증] AGI 개발기관은 이사회 구성원, 고위 임원, 핵심 직원을 고용하거나 임명하기 전에 철저한 신원 및 이력 검증을 수행해야 한다.
Board risk committee*	[이사회 내 위험위원회] AGI 개발기관은 이사회 내에 상설위험위원회, 즉 개발기관의 위험관리 관행을 상시 감독하는 위원회를 두어야 한다.
Bug bounty programs	[버그바운티 프로그램] AGI 개발기관은 버그바운티 프로그램을 운영해야 한다. 즉, AGI 기관은 알려지지 않은 취약점과 위험한 능력을 보고한 자를 인정하고 보상하는 버그바운티 프로그램을 운영해야 한다.
Chief risk officer*	[최고위험관리자] AGI 개발기관은 위험관리를 책임지는 고위 임원인 최고위험관리자(CRO)를 두어야 한다.
Dangerous capabilities evaluations	[위해 가능성 평가] AGI 개발기관은 자사 모델의 위해 가능성을 평가하여야 한다.
Dual control*	[이중 통제] 모델 개발 및 배포에 관한 중대한 결정은 최소 두 명이 함께 내려야 한다.
Emergency response plan	[비상 대응 계획] AGI 개발기관은 비상 상황 대응 계획을 수립하고 정기적으로 연습해야 한다. 이는 시스템 종료, 출력 무효화, 접근 제한 등을 포함할 수 있다.
Enterprise risk management	[기관 전체의 위험관리] AGI 개발기관은 기관 전체의 위험관리 프레임워크를 시행해야 하고, 이 프레임워크는 AG 맥락에 맞게 조정되어야 하며, 개발기관이 사회에 미치는 영향에 중점을 두어야 한다.
Gradual scaling	[점진적 규모 확대] AGI 개발기관은 최대 규모 모델의 훈련에 사용되는 컴퓨팅 자원과 데이터의 양을 점진적으로만 증가시켜야 한다.
Increasing levels of external scrutiny	[외부 감독 강화] AGI 개발기관은 모델의 능력에 비례하여 외부 감독 수준을 높여야 한다.
Industry sharing of security information*	[보안 정보 공유] AGI 개발기관은 위협 인텔리전스와 보안사고 정보를 서로 공유해야 한다.
Inter-lab scrutiny*	[개발기관 간 검토] AGI 개발기관은 타 개발기관의 연구자들이 배포 전에 강력한 모델을 검토할 수 있도록 허용해야 한다.
Internal audit	[내부 감사] AGI 개발기관은 연구소의 위험관리 관행의 효과성을 평가하는 내부 감사팀을 두어야 한다. 이 팀은 조직적으로 고위 경영진으로부터 독립적이어야 하며 이사회에 직접 보고해야 한다.
Internal review before publication	[연구 발표 전 내부 검토] AGI 개발기관은 연구 결과를 발표하기 전에 잠재적 위험을 평가하는 내부 검토를 수행해야 한다.
KYC screening*	[고객확인심사] AGI 개발기관은 강력한 모델 사용 권한을 부여하기 전에 고객확인심사를 수행해야 한다.
Military-grade information security	[군용 등급 정보 보안] AGI 개발기관의 정보 보안은 모델의 능력에 비례해야 하며, 궁극적으로는 정보기관의 수준과 동일하거나 그 이상이어야 한다.
Model containment	[모델 격리] AGI 개발기관은 위해 가능성이 매우 높은 능력을 보유한 모델을 격리하여야 한다.
Monitor systems and their uses	[시스템 및 사용 모니터링] AGI 개발기관은 배포된 시스템의 사용 방식과 사회적 영향을 면밀히 모니터링해야 한다.
No unsafe open-sourcing	[안전하지 않은 경우 오픈소스화 금지] AGI 개발기관은 충분히 안전하다고 입증할 수 없는 한 강력한 모델을 오픈소스로 공개하지 않아야 한다.
Notify a state actor before deployment	[배포 전 정부 기관 통보] AGI 개발기관은 강력한 모델을 배포하기 전에 관계 당국에 통보해야 한다.
Notify affected parties*	[영향받는 당사자에게 통보] AGI 개발기관은 강력한 모델로 인해 부정적인 영향을 받을 수 있는 당사자들에게 배포 전에 통보해야 한다.

[참고] 50개 관행 – 알파벳 순 정렬

영문 축약	[한글 축약] 관행
Notify other labs*	[타 개발기관에 통보] AGI 개발기관은 강력한 모델을 배포하기 전에 타 개발기관에 통보해야 한다.
Pausing training of dangerous models	[위험 모델 개발 중단] AGI 개발기관은 충분히 위험한 능력이 감지되면 개발 과정을 중단해야 한다.
Post-deployment evaluations*	[배포 후 평가] AGI 개발기관은 배포 후에도 모델의 기능과 사용 방법에 대한 새로운 정보를 고려한 모델의 위해 가능성을 지속적으로 평가해야 한다.
Pre-deployment risk assessment	[배포전 위험평가] AGI 개발기관은 강력한 모델을 배포하기 전에 광범위한 위험식별, 분석 및 평가를 수행해야 한다.
Pre-registration of large training runs	[대규모 모델 훈련 사전 등록] AGI 개발기관은 일정 규모 이상의 모델 훈련을 관계 당국에 사전 등록해야 한다.
Pre-training risk assessment	[사전 훈련 위험평가] AGI 개발기관은 강력한 모델을 훈련시키기 전에 위험평가를 수행해야 한다.
Protection against espionage*	[산업 스파이 방지] AGI 개발기관은 간첩 또는 산업 스파이로부터의 위험에 대비한 적절한 조치를 취해야 한다.
Publish alignment strategy*	[정렬 전략 공개] AGI 개발기관은 시스템의 안전성과 정렬을 보장하기 위한 전략을 공개해야 한다.
Publish results of external scrutiny*	[외부 감사 결과 공개] AGI 개발기관은 외부 감사 결과 또는 요약물을 공개해야 한다. 단, 이것이 독점정보를 과도하게 노출하거나 중대한 위험을 초래하지 않는 경우에 한한다.
Publish results of internal risk assessments*	[내부 위험평가 결과 공개] AGI 개발기관은 내부 위험평가 결과 요약물을 공개해야 하고, 기존 위험을 감수하는 이유를 포함해야 한다. 단, 독점정보를 과도하게 노출, 중대한 위험을 초래하지 않을 경우로 한정된다.
Publish views about AGI risk	[AGI 위험에 대한 견해 공개] AGI 개발기관은 AGI의 위험과 이점에 대한 견해와 개발 과정에서 감수할 수 있는 위험 수준에 대해 공개적으로 밝혀야 한다.
Red teaming	[외부 레드티밍] AGI 개발기관은 강력한 모델을 배포하기 전에 외부 레드팀을 통한 취약점 분석을 수행해야 한다.
Report safety incidents	[안전사고 보고] AGI 개발기관은 사고나, 아차사고를 관련 정부 기관과 다른 AGI 개발기관에 보고해야 한다.
Researcher model access	[독립연구자의 모델 접근] AGI 개발기관은 독립 연구자들에게 배포된 모델에 대한 API 접근 권한을 제공해야 한다.
Safety restrictions	[안전 제한 조치] AGI 개발기관은 강력한 모델 배포 후 적절한 안전 제한 조치를 취하여야 한다.
Safety vs. capabilities	[안전 vs. 능력] AGI 개발기관 직원의 상당 부분은 모델의 능력 향상보다는 안전성과 정렬 개선에 집중해야 한다.
Security incident response plan*	[보안 사고 대응 계획] AGI 개발기관은 보안 사고에 대한 대응 계획을 수립해야 한다.
Security standards	[보안 표준 준수] AGI 개발기관은 정보 보안 표준을 준수해야 하고, 이 표준들은 AGI 맥락에 맞게 조정되어야 한다.
Staged deployment	[단계적 배포] AGI 개발기관은 강력한 모델을 단계적으로 배포해야 한다. 소수의 응용 프로그램과 사용자로 시작하여 모델의 안전성에 대한 확신이 커짐에 따라 점진적으로 확대해야 한다.
Statement about governance structure*	[거버넌스 구조 공개] AGI 개발기관은 모델 개발 및 배포와 관련된 중요 결정을 어떻게 내리는지에 대한 거버넌스 구조를 공개해야 한다.
Third-party governance audits*	[제3자 거버넌스 감사] AGI 개발기관은 자체 거버넌스 구조에 대한 제3자 감사를 의뢰해야 한다.
Third-party model audits	[제3자 모델 감사] AGI 개발기관은 강력한 모델을 배포하기 전에 제3자 감사를 의뢰해야 한다.
Tracking model weights*	[모델 가중치 추적] AGI 개발기관은 강력한 모델의 가중치 복사본을 모두 추적할 수 있는 시스템을 갖춰야 한다.
Treat internal deployments similar to external deployments*	[외부 배포에 준하는 내부 배포] AGI 개발기관은 내부 배포를 외부 배포와 유사하게 처리하여야 하고, 특히 AGI 개발기관은 배포 전 위험평가를 수행해야 한다.
Treat updates similarly to new models	[새 모델과 유사하게 업데이트 모델 위험평가] AGI 개발기관은 배포된 모델의 중요한 업데이트(예: 추가 파인튜닝)를 초기 개발 및 배포와 유사하게 처리해야 한다. 특히 배포 전 위험평가를 다시 수행해야 한다.

저자 소개

정선화 ETRI ICT전략연구소 기술정책연구본부 기술경제연구실 책임연구원
e-mail: sh-jeong@etri.re.kr / Tel. 042-860-6511

AGI 안전 및 거버넌스 관행에 관한 국내외 전문가 인식 비교

발행인 한 성 수
발행처 한국전자통신연구원 ICT전략연구소
발행일 2024년 12월 31일