NAVER AI Safety Framework (ASF)

최초 제정일: 2024년 6월 17일
문서 관리자: NAVER AI Safety Center

NAVER

목차

ASF의 관점

네이버는 AI의 개발과 이용에 있어 인간 중심의 가치를 최우선으로 생각하며, AI를 사용자를 위한 일상의 도구로 발전시켜 나가고 있습니다.

2021년 네이버 AI 윤리 준칙을 발표한 이후 사람을 위한 AI 개발 및 사용을 위해 다양한 노력을 기울여 온 데 이어, AI 시스템(AI 모델과 AI 서비스를 포함)과 관련된 위험을 예방하기 위해 글로벌 흐름에 맞춰 NAVER ASF(AI Safety Framework)를 구체화해 나가고자 합니다.

NAVER ASF는 AI Safety와 관련해 사회에서 우려하고 있는 위험에 대응하기 위한 체계입니다. 이를 통해, 네이버는 AI 시스템의 개발 및 배포 프로세스의 전 단계에서 관련된 위험을 인식, 평가 및 관리합니다. 또한, AI 기술 발전에 따른 AI Safety 관련 글로벌 논의 흐름에 맞춰, ASF를 지속적으로 개선하며 업데이트해 나갈 것입니다.

네이버는 다양성을 통해 연결이 더 큰 의미를 가질 수 있도록 기술과 서비스를 구현해 왔고, 그 과정에서 사용자에게 다채로운 기회와 가능성을 열어 왔습니다. 네이버는 AI를 개발하는 데 있어서도 사람을 위한 AI라는 가치와 함께 다양성도 조화롭게 고려할 필요가 있다고 믿으며, AI Safety에 있어서도 글로벌 AI Safety 움직임에 발맞추는 한편 각 지역의 사회기술적 맥락(socio-technical context)을 고려해 접근하는 것이 중요하다고 생각합니다.

위험 인식

AI Safety와 관련해 사회에서 우려하고 있는 위험은 크게 통제력 상실 위험과 악용 위험, 두 가지로 정리해 볼 수 있습니다.

AI 시스템이 지속적으로 발전하면서 통제력을 상실하지 않을지 우려하는 것과 AI 시스템이 위험을 초래할 수 있는 영역에서 악용되거나 오남용될 위험이 있지 않을까 하는 부분입니다. 그리고 이 밖에도 현재 AI 기술이 가지고 있는 기술적 한계점 역시 사회에서 지속적으로 논의되고 있습니다.

NAVER ASF에서 정의하고 있는 통제력 상실 위험은 미래에 인간이 AI 시스템에 영향을 미치지 못하게 되는 중대한 위험을 말합니다(severe disempowerment of the human species). 그렇기 때문에 통제력 상실 위험은 개념 정의 상 현재 존재하는 AI를 활용한 자동화를 포함하는 것은 아닙니다. 해당 위험은 AI 시스템의 성능이 개선됨에 따라 지속적으로 증가하는 유형의 위험은 아니지만, AI 시스템이 기술적으로 고도화되는 경우 통제력 상실 위험이 발생할 수도 있다는 시각이 있습니다. 네이버는 이러한 사회의 우려를 고려해 해당 위험에 대해 일정한 기준을 갖고 지속적으로 살펴보는 것이 필요하다고 생각하고 있습니다.

NAVER ASF에서 정의하고 있는 악용 위험은 AI 시스템의 기술적 고도화와 무관하게 생화학 물질 개발 영역과 같이 사회적으로 우려되는 영역에 활용되거나, AI 시스템의 목적과 달리 악용될 가능성이 있는 위험을 말합니다. 이러한 위험을 완화하기 위해서는 AI 시스템이 사회적으로 우려되는 영역에 활용되지 않도록 하거나, AI 시스템의 목적과 달리 악용될 가능성을 줄일 수 있도록 안전 조치를 할 필요성이 있습니다. 네이버는 현재까지 기술적, 정책적인 조치를 포함해 다양한 조치를 진행해왔고, 앞으로도 사용자의 안전을 위해 다양한 조치를 진행할 계획입니다.

평가 및 관리

NAVER ASF는 네이버 AI 윤리 준칙을 준수하는 네이버 구성원이 산업 현장에서 AI 시스템을 개발하고 배포하는 과정에서 AI Safety를 구체적으로 실천하기 위한 체계입니다. 이를 통해 네이버는 사회에서 우려하고 있는 AI에 대한 통제력 상실 위험과 악용 위험에 대처하고자 합니다.

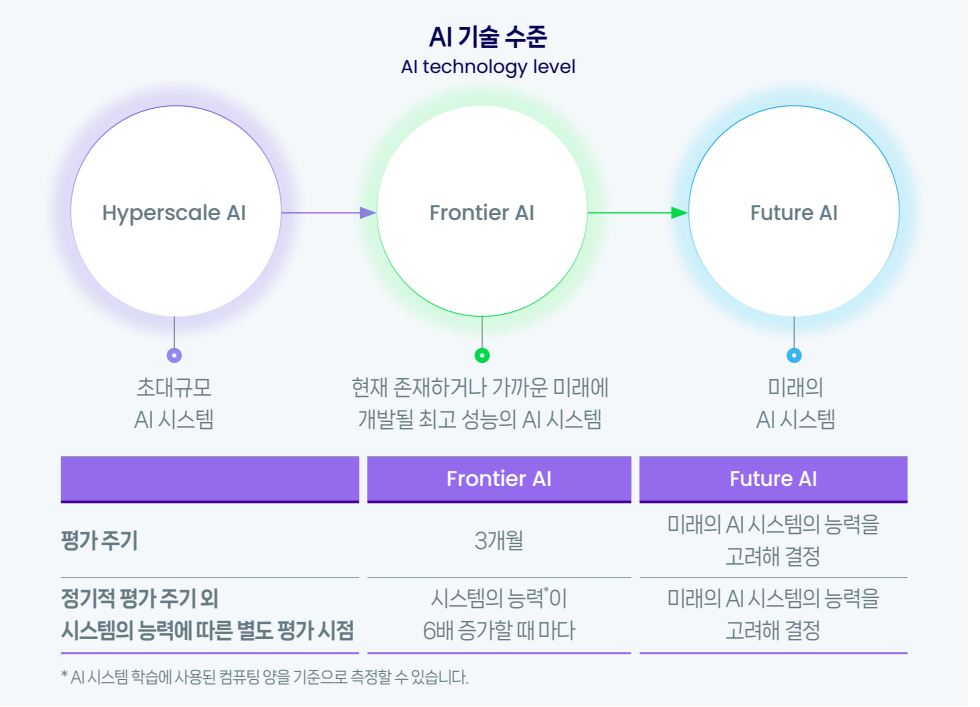
통제력 상실 위험에 대해서는 AI 위험 평가 스케일을 통해 대응하고, 악용 위험에 대해서는 AI 위험 평가 매트릭스를 통해 위험을 관리합니다. 특히, 이를 실천하는 과정에서 사회기술적 맥락을 고려하기 위해 노력하고 있습니다. 사회기술적 맥락을 고려한 AI 시스템을 만들기 위해서는 AI 시스템을 특정 문화와 사회에 맞는 데이터셋으로 학습하고, 평가하는 것이 중요합니다.

평가 및 관리

A. AI 위험 평가 스케일

AI 위험 평가 스케일은 통제력 상실 위험에 대비하기 위한 체계로, AI 시스템의 능력이 발전함에 따라 통제력 상실 위험이 발생할 수 있는지 여부를 AI 시스템 개선에 맞춰 주기적으로 또는 능력에 따라 필요한 경우 살펴보고 위험을 인식하고 평가, 관리합니다.

AI 위험 평가 스케일에서는 AI 기술 수준을 Hyperscale AI, Frontier AI, Future AI 세 가지로 나누며, 이 가운데 Frontier AI를 중심으로 평가를 진행합니다. Future AI는 미래의 AI 시스템으로, 해당 AI 시스템의 능력을 고려해 평가 주기와 평가 시점을 추후 결정하려고 합니다.



	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	미래의 AI 시스템의 능력을 고려해 결정

* AI 시스템 학습에 사용된 컴퓨팅 양을 기준으로 측정할 수 있습니다.

Frontier AI는 현재 존재하는 또는 가까운 미래에 개발될 최고 성능의 AI 시스템입니다. 3개월을 평가 주기로 설정하여 사회적으로 우려되는 통제력 상실 위험에 대비하고자 합니다. 일반적으로는 3개월마다 주기적으로 평가하지만, 만약 시스템의 능력이 기존보다 6배 증가했다고 판단되면 해당 시점에 별도로 평가를 진행하려고 합니다. AI 시스템의 능력은 보통 모델의 규모가 커질수록 증가하므로, 학습에 사용된 대략적인 컴퓨팅양을 기준으로 삼을 수 있습니다.

B. AI 위험 평가 매트릭스

사회에서 우려하는 악용 위험에 대비하기 위해서 네이버는 AI 위험 평가 매트릭스를 적용하여 위험을 관리하고 있습니다.

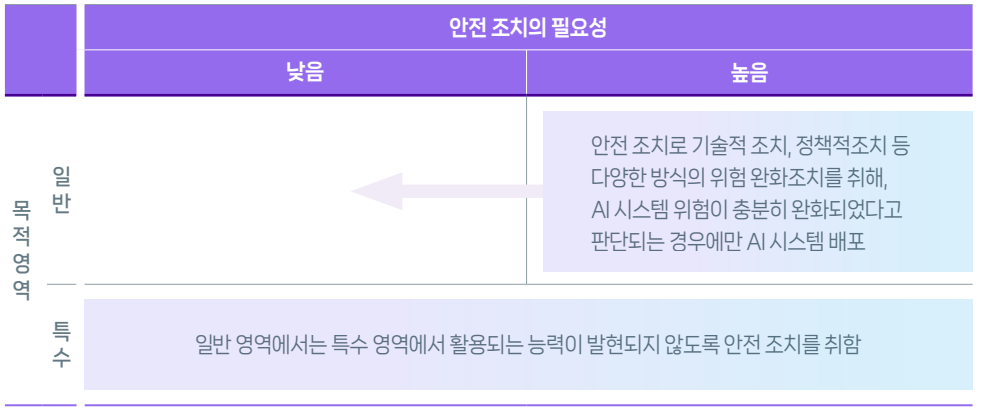
	목적 영역	안전 조치의 필요성
위험 평가	AI 시스템이 활용되는 영역을 고려해 특수한 영역에 사용될 목적이라면 해당 영역에서 AI 시스템 위험이 발생할 수 있는지 평가함.	전체 라이프사이클에 맞춰 기업 내 다양한 부서와 협업하여 AI 시스템의 위험을 인식하고, 해당 위험이 발생할 수 있는지 평가함.
위험 관리 방법	특수 영역에서 사용될 목적인 경우, 해당 AI 시스템은 특별한 자격이 있는 사용자에게만 제공될 수 있도록 안전 조치해 위험을 관리함. 목적 영역 구분을 위해, 특수 영역이 아닌 일반 영역에서는 특수 영역에서 활용되는 능력이 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

평가 및 관리

AI 위험 평가 매트릭스는 AI 시스템의 목적 영역과 안전 조치의 필요성이라는 두 가지의 기준을 토대로 AI 시스템 위험이 발생할 수 있는지 여부를 AI 시스템의 전체 라이프사이클에 맞춰 살펴보고 위험을 인식, 평가, 관리합니다.

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

두 가지 기준을 종합하여 인식, 평가된 AI 시스템의 위험에 따라 적절한 조치를 시행합니다. 안전 조치를 통해 충분히 위험이 완화되었다고 판단되는 경우에만 AI 시스템을 배포하고, 배포 이후에도 안전성 모니터링을 지속적으로 시행하여 AI 시스템 위험을 관리합니다. 이론적으로는 특수 영역이면서도 안전 조치의 필요성이 높은 경우가 존재할 수 있는데, 이런 상황에서는 AI 시스템을 배포하지 않도록 하여 AI 시스템 위험을 관리하고자 합니다.



일반 영역에서는 특수 영역에서 활용되는 능력이 발현되지 않도록 안전 조치를 취해 위험을 완화합니다. 안전 조치가 필요하다고 판단되는 경우에는 기술·정책 측면 등 다방면으로 조치를 취해 AI 시스템 위험을 완화하고, 충분히 위험이 감소했다고 판단될 때만 AI 시스템을 배포해서 위험을 관리합니다.

실제로 네이버는 네이버 AI 윤리 준칙 과 논문 연구를 바탕으로 다양한 안전 조치를 적용하고 있습니다. 이를 위해, 네이버는 학계, 산업 등 다양한 이해관계자와 함께 한국 문화와 사회에 맞는 한국어 데이터셋을 새롭게 구축하고, 기존의 영미권 문화를 바탕으로 만들어진 벤치마크 데이터에 한국의 특성을 반영하는 연구를 진행했습니다.

연구를 통해 구축된 데이터셋(SQuARe*, KoSBi, KoBBQ* 등)은 하이퍼클로바X에 활용되었습니다. 네이버는 이러한 연구 결과 일부를 누구나 활용할 수 있도록 데이터를 공개하고 있으며 , 다양한 이해관계자와의 협업을 통해 안전성 연구를 지속적으로 진행하고 있습니다.

* SQuARe, KoSBi: https://github.com/naver-ai/korean-safety-benchmarks
* KoBBQ: https://huggingface.co/datasets/naver-ai/kobbq

거버넌스

네이버는 NAVER ASF를 실천하기 위한 관리구조를 갖춰 나가고 있습니다. 네이버는 관리 구조에 따라 다양한 분야에 전문성을 보유한 구성원들의 협업을 통해 네이버가 개발하는 AI 시스템의 위험을 인식, 평가, 관리하고자 합니다.

네이버의 AI Safety 거버넌스는 아래와 같습니다.

- Future AI Center는 네이버 내 다양한 부서가 참여하는, AI 시스템 위험에 대한 실무적인 논의 기구입니다.
- 리스크관리워킹그룹은 실무적으로 논의된 AI 시스템 위험에 대해 이사회에 보고할 사항을 판단하는 기구입니다.
- 이사회(리스크관리위원회)는 AI 시스템 위험에 대한 최종적인 의사결정 기구입니다.

외부와의 협업

네이버는 기업 외부의 이해관계자들과 함께 AI 기술과 서비스에 대한 우려 사항을 해결하기 위해 협업하고 있습니다. 대표적으로 서울대, 카이스트와는 AI 기술 분야에서, 서울대 인공지능 정책 이니셔티브를 통해서는 AI 정책 분야에서 협력하고 있습니다.

전문가와의 협업과 함께 사용자와의 협력도 확대해 나가고 있습니다. 작게는 AI 서비스를 통해 사용자 피드백을 받고 있으며, 사람을 위한 AI 활용 가이드의 사례처럼 사용자와 AI 서비스를 어떻게 활용할 것인지에 대한 협력도 진행하고 있습니다.

외부 이해관계자와의 협업 사례로는 2024년 4월, 정부 기관, 생성형 AI 기업, 여러 분야의 참가자들과 함께 진행한 '생성형 AI 레드팀 챌린지'가 있습니다. 레드팀 챌린지는 다양한 참가자들이 탈옥, 편견·차별, 인권침해, 사이버 공격, 불법 콘텐츠, 잘못된 정보, 비일관성 등 7개 주제를 대상으로 잠재적 위험과 취약점을 찾는 방식으로 진행됐습니다. 참가자들은 유해한 콘텐츠를 유도하거나 특정 사회적 집단에 대한 고정관념이나 편견에 근거한 부정적인 응답을 유도하며 생성형 AI 모델의 취약점을 발굴했습니다.

앞으로의 노력

네이버가 개발하고 이용하는 AI는 사용자를 위한 일상의 도구입니다. 네이버는 AI의 개발과 이용에 있어 인간 중심의 가치를 최우선으로 생각하며, AI 시스템과 관련된 위험을 예방하기 위해 NAVER ASF를 구현하고 실천해 나가겠습니다.

네이버는 사용자의 일상에 편리함을 더하기 위해 기술을 개발해 왔고, AI 역시 일상의 도구로 활용될 수 있도록 발전시켜 나가고 있습니다. 세상의 다른 모든 것처럼 AI 시스템이 완벽할 수 없다는 점을 인식하고, 사람을 위해 사용될 수 있도록 AI 시스템을 필요한 시점에 맞춰 지속적으로 살펴보겠습니다.

AI 기술 발전에 따른 AI Safety 관련 논의 흐름에 맞춰, 네이버는 NAVER ASF를 지속적으로 개선해 나가고자 합니다. 특히, 다음 사항에 대한 고민을 통해 NAVER ASF를 더욱 구체화할 수 있도록 노력하겠습니다.

앞으로의 노력

안전한 소버린 AI 공동개발

네이버는 세계에서 세 번째로 자국어 중심 초대규모 소버린 언어모델과 산업생태계를 만들어 가면서, 문화적, 지정학적 상황과 지역적 이해가 AI의 성능과 Safety에도 영향을 미칠 수 있다는 점을 알게 됐습니다. 네이버는 이같은 경험을 글로벌 커뮤니티에 공유하며, 각국의 사회기술적 맥락(socio-technical context)을 반영한 안전한 소버린 AI를 공동으로 개발해 나가도록 하겠습니다.

공동 개발 과정에서 AI Safety와 관련해 다음과 같은 과정들을 거칠 것으로 예상됩니다:

- 각 문화권의 니즈에 부합하는 활용사례(use case)의 발굴과 리스크(risk) 식별
- 각 문화권에 특유한 리스크를 제대로 측정할 수 있는 벤치마크(benchmark)의 고도화
- 벤치마크에 상응하는 다양한 도구들의 개발과 보급

ASF 통한 AI 정책 고도화

네이버는 AI 시스템의 라이프 사이클, 특히 서비스 배포 전 시점에 기업 내 다양한 부서가 협업해 발생할 수 있는 위험을 인식, 평가, 관리하는 정책을 설계해 왔습니다. AI 윤리 자문프로세스(CHEC: Consultation on Human-Centered AI's Ethical Considerations) 등 기존 AI 정책에 NAVER ASF를 반영해 고도화시켜 나가겠습니다.

AI Safety 관리구조 구체화

네이버는 ASF 구현과 실천을 위해 AI Safety 관리구조를 구체화하고 개선해 나가겠습니다. 또한, 기업 외부의 전문가, 사용자 등과도 협력해 나가며 AI 안전성을 높여 나가겠습니다.

글로벌 AI Safety 움직임 동참

AI 기술 발전에 따른 AI Safety 관련 글로벌 논의 흐름을 반영해 ASF를 지속적으로 개선하며 업데이트해 나가도록 하겠습니다. AI Safety와 관련해 국내외 AI 안전 연구소 네트워크, Frontier Model Forum, UN, OECD 등과 긴밀히 협력 하겠습니다.


NAVER