NAVER AI Safety Framework (ASF)

Initial Publication Date: June 17, 2024
Managed by: NAVER AI Safety Center

NAVER

Table of Contents

Our AI Safety Framework

NAVER takes a human-centric approach to developing AI, and our aim is to help people benefit from AI by turning this technology into a daily tool.

Since introducing NAVER's AI Principles in 2021, human-centered AI development has always been the focus of our efforts. By building on our AI Safety Framework, we hope to proactively address potential harms related to AI systems.

Our AI Safety Framework is designed to address societal concerns around AI safety. We identify, assess, and manage risks at all stages of AI systems operations, from development to deployment. Technological advances in AI have accelerated the shift toward safer AI globally, and our AI safety framework, too, will be continuously updated to keep up with changing environments.

NAVER has always endorsed diversity when launching new technologies and services to make connections genuinely worthwhile. We believe that developing people-centered AI must go hand in hand with respecting diversity. While being part of the international community that works toward safer AI, we're also dedicated to approaching AI within the socio-technical context from the perspective of a local society.

Risk Awareness

The potential harms of AI that many people voice concern over broadly fall into one of two categories: “loss of control” and “misuse” risks.

The former concerns the fear of losing control over AI systems as they become more sophisticated, while the latter refers to the possibility of people deliberately manipulating these systems to catastrophic effect. AI’s technological limitations are also a key point in discussions about trust and safety.

NAVER’s AI Safety Framework defines the first category of risk as AI systems causing severe disempowerment of the human species. By this definition, this loss of control risk goes far beyond the implications of current AI-enabled automation, which stems from the concern that AI systems could spiral out of human control at the pace they are advancing. At NAVER, we take this risk seriously as we continually apply our standards to look for signs of alarm.

Our AI Safety Framework describes the second risk category as misusing AI systems to develop hazardous biochemical weapons or otherwise use them against their original purpose. To mitigate such risks, we have to place appropriate safeguards around AI technology. NAVER has taken a wide range of technological and policy actions so far and will continue to work toward achieving AI safety.

Risk Assessment and Management

Our AI Safety Framework outlines specific actions for everyone at NAVER working in the field, for those who develop and deploy AI systems and are committed to our AI Principles. With this framework, we hope to address the “loss of control” and “misuse” risks.

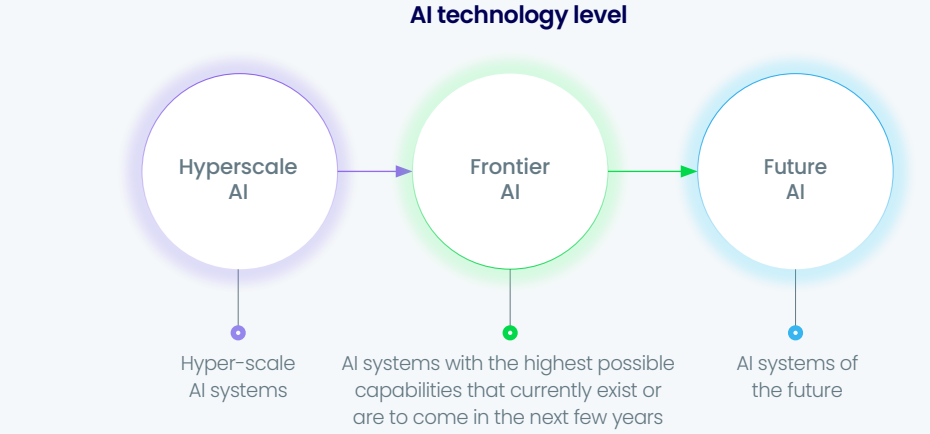
We use the risk assessment scale for the former category and the risk assessment matrix for the latter to manage potential risks. When building AI systems, we try to do so in the socio-technical context, which means training and evaluating language models with local datasets within the proper cultural and societal context.

Risk Assessment and Management

A. AI Risk Assessment Scale

The risk assessment scale examines risks in the “loss of control” category to see whether they are positively correlated with the advancement of AI systems. LLMs should be subject to periodic reviews or assessed whenever major performance improvements are made.

Depending on their technological level, AIs can be divided into three types: hyper-scale, frontier, and future AI. For the purpose of our evaluation, we’ll focus on frontier AI. Future AI refers to AI systems that have yet to arrive and can only be evaluated once their capabilities become clearer.



	Hyperscale AI	Frontier AI	Future AI
Evaluation cycle	Non-existent	Every 3 months, or when performance increases by 6x	To be determined laterdepending on their future capabilities

* Can be estimated based on the amount of computing used to train the AI system

Frontier AI possesses the top capabilities that are available today or will be soon in the near future. Our goal is to have AI systems evaluated quarterly to mitigate loss of control risks, but when performance is seen to have increased six times, they will be assessed even before the three-month term is up. Because the performance of AI systems usually increases as their size gets bigger, the amount of computing can serve as an indicator when measuring capabilities.

B. AI Risk Assessment Matrix

For the other “misuse” risk category, we use a risk assessment matrix to manage risks.

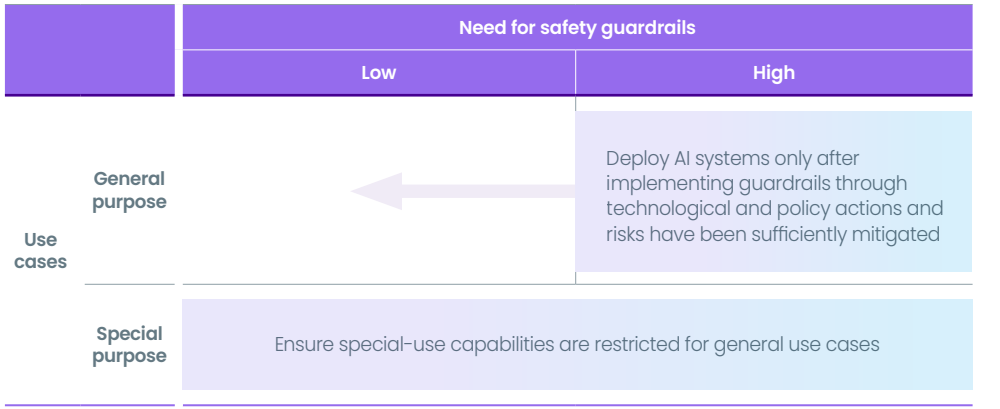
	Use cases	Need for safety guardrails
Assessing risks	Determine whether an AI system designed to serve a certain purpose can cause potential harm in special use cases	Collaborate with different teams to identify and calculate the probability of risks across the entire lifecycle
Managing risks	For special use cases, make AI systems available only to authorized users For general use cases, build guardrails by restricting special-use capabilities	Delay deploying AI systems until risks are mitigated and appropriate technological and policy actions have been taken
Examples	Special use case: biochemical weaponization	Technological measure: AI safety updates on HyperCLOVA X models

Risk Assessment and Management

The AI risk assessment matrix is used to identify, evaluate, and manage risks during an AI system's entire lifecycle on two grounds: its purpose or use case and the need for safety guardrails.

		Need for safety guardrails	
		Low	High
Use cases	General purpose	**Low risk** Deploy AI systems but perform monitoring afterward to manage risks	**Risk identified** Withhold deployment until additional safety measures are taken
	Special purpose	**Risk identified** Open AI systems only to authorized users to mitigate risks	**High risk** Do not deploy AI systems

Once AI systems are evaluated and their risks identified according to the two standards, we must implement appropriate guardrails around them. We should only deploy AI systems if those safeguards have proven effective in mitigating risks and keep an eye on the systems even after deployment through continuous monitoring. In theory, there may be cases where AI systems are used for special purposes and require safety guardrails in place, in which case AI systems should not be deployed.



Keep special-use capabilities off limits for general purposes by placing appropriate safeguards. If safety guardrails are deemed necessary, take technological and policy actions to reduce risks and deploy only when risks are sufficiently mitigated.

We draw from our AI Principles and research studies to implement guardrails around AI models. Our partnerships with academia, the tech industry, and other stakeholders have led to meaningful research in building local datasets specialized to Korean culture and society and creating benchmarks for evaluating Korean-centric models.

We apply SQuARe*, KoSBi*, KoBBQ** datasets built from our research to our HyperCLOVA X models. We also make parts of these studies available to everyone as we continue to collaborate with various stakeholders in conducting safety research.

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

AI Governance

NAVER's AI Safety Framework is our initiative to achieve AI governance. Under our governance, we foster collaboration between cross-functional teams to identify, evaluate, and manage risks when developing AI systems.

NAVER's AI governance includes:

- The Future AI Center, which brings together different teams for discussions on the potential risks of AI systems at the field level
- The risk management working group whose role is to determine which of these issues to raise to the board
- The board (or the risk management committee) that makes the final decisions on the matter

External Collaborations

We work with external stakeholders to take on challenges surrounding safe AI technologies and services.

Building trust in AI is a collective effort, which is why we partner with top universities like Seoul National University (SNU) and Korea Advanced Institute of Science & Technology (KAIST) on the technology front and participate in the SNU AI Policy Initiative on the policy front.

But it's not just the experts we're collaborating with. We also work with our users, improving AI through their feedback and writing guides on Using AI for People.

In April this year, we held our first Generative AI Red Teaming Challenge. Participants screened LLMs for signs of harmful content across seven domains—human rights violation, disinformation, inconsistency, cyberattacks, bias and discrimination, illegal content, and jailbreaking—exposing vulnerabilities by triggering the models to generate unethical responses that were biased or discriminatory.

Looking Ahead

The AI that NAVER develops and uses is an everyday tool for our users. We put human-centered values first in how AI is developed and used, and we will continue to implement and practice NAVER ASF to prevent risks associated with AI systems.

NAVER has always developed technology to bring greater convenience to people's daily lives, and we are advancing AI so that it, too, can serve as a tool for everyday use. Recognizing that AI systems—like everything else in the world—cannot be perfect, we will continue to review our AI systems whenever necessary to ensure they are used for the benefit of people.

In step with the evolving global discussion on AI safety that accompanies technological progress, we intend to keep improving NAVER ASF. In particular, we will work to make NAVER ASF more concrete by giving careful consideration to the following.

앞으로의 노력

Co-developing Safe Sovereign AI

As the third in the world to build a hyperscale sovereign language model centered on the Korean language—along with the industrial ecosystem around it—NAVER has come to understand that cultural and geopolitical circumstances and local perspectives can affect both the performance and the safety of AI. We will share this experience with the global community and work jointly to develop safe sovereign AI that reflects the socio-technical context of each country.

In the course of such co-development, we expect the following AI safety-related work to take place:

- Identifying use cases that meet the needs of each cultural sphere, along with the risks they carry
- Advancing benchmarks capable of properly measuring risks specific to each cultural sphere
- Developing and distributing a range of tools that correspond to those benchmarks

Advancing AI Policy through the ASF

NAVER has designed policies through which teams across the company collaborate to identify, assess, and manage risks that may arise over the AI system lifecycle, particularly before a service is deployed. We will further advance our existing AI policies—including CHEC (Consultation on Human-Centered AI's Ethical Considerations), our AI ethics advisory process—by reflecting NAVER ASF in them.

Building Out Our AI Safety Governance Structure

To implement and practice the ASF, we will give concrete shape to our AI safety governance structure and continue to improve it. We will also collaborate with external experts, users, and other stakeholders to raise the level of AI safety.

Joining the Global AI Safety Movement

We will continue to improve and update the ASF to reflect the flow of global discussion on AI safety accompanying advances in AI technology. On AI safety, we will work closely with domestic and international AI safety institute networks, the Frontier Model Forum, the UN, and the OECD, among others.


NAVER