AI Safety Framework 2.0

NAVER ASF 2.0

Initial Publication Date: July 7, 2026
Managed by: NAVER AI Safety Center

NAVER

Table of Contents

1. Preface

At NAVER, human-centered values are at the core of how we develop and use AI. We are advancing cutting-edge AI technology to make it an accessible, everyday tool for everyone.

After publishing the “NAVER AI Ethics Principles” in 2021, we established a framework for identifying and managing the risks of frontier AI models through “NAVER ASF (AI Safety Framework) Beta” in 2024. Since then, we have continuously refined NAVER ASF in step with global developments and the evolving domestic policy landscape.

As the technology, services, policies, and regulations surrounding AI continue to evolve, AI Safety is no longer just about making a single model safe. It now extends to how we safely design and operate services—built by combining diverse AI models—that serve tens of millions of people.

Through “NAVER ASF 2.0”, we will continue to refine and advance our service-centered AI Safety governance, giving concrete shape to AI Safety across our On-Service AI.

2. The Direction of ASF 2.0

NAVER ASF is our AI Safety framework for addressing the AI risks that concern society. Through the "NAVER AI Ethics Principles" and "NAVER ASF Beta", we have worked to proactively prevent and mitigate the risks of AI systems.

As AI technology is embedded in real services and becomes part of users' daily lives, AI Safety is moving beyond theoretical discussion to directly tackle real-world challenges. These challenges arise within the unique context of each individual service. It is no longer enough to consider the safety of an AI model alone—we must also account for the environment in which a service operates, how it interacts with users, and the situations that may arise along the way.

To respond to these changes, NAVER has organized its AI Safety Framework under the direction of On-Service AI. Through this framework, we will examine from multiple perspectives how AI services may be used, and the impact they may have on users, the broader public, and the AI ecosystem as a whole—and work to address the issues that emerge.

The Evolving Landscape of AI Safety

The Advancement of On-Service AI NAVER is advancing its On-Service AI strategy, integrating AI technology across its services to expand the user experience. We have launched AI Briefing, AI Shopping Agent, and AI Tab, and plan to introduce a range of vertical agent services. Because the types of risk and required safety levels can vary depending on a service's domain and scope—even for the same AI model—safety management needs to extend from the model to the service.

The Spread of Multi-Model Environments At the time of "NAVER ASF Beta", generative AI services typically relied on a single large language model to handle all requests. More recently, multi-model environments—combining internal and external models, large and small, according to their purpose—have become increasingly common. In these environments, safety reviews must go beyond individual models to also examine risks that may emerge from how models are combined and the context in which they are used within a service.

Changes in the Policy and Regulatory Environment On January 22, 2026, the Framework Act on the Development of Artificial Intelligence and the Creation of a Foundation for Trust (hereinafter, the "AI Basic Act") came into effect. In its approach to AI Safety, NAVER keeps pace with both the domestic policy environment and global developments. We are advancing our AI Safety processes to concretely define the real, specific challenges facing AI services within their sociotechnical context—and to identify, assess, manage, and improve upon the associated risks.

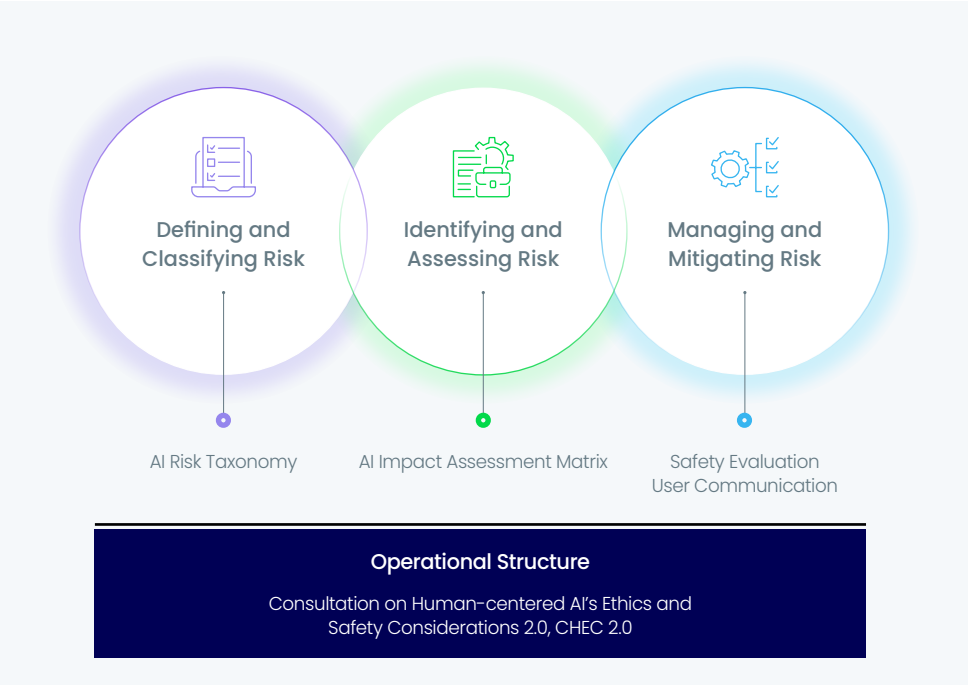
ASF 2.0 is the service-centered AI Safety Framework we have established to respond to these shifts—the advancement of On-Service AI, the spread of multi-model environments, and changes in the policy and regulatory environment.

2. The Direction of ASF 2.0

Overview of ASF 2.0

ASF 2.0 is built on three pillars—defining and classifying risk, identifying and assessing risk, and managing and mitigating risk—along with an operational structure that ensures these pillars work as intended in practice. The three pillars define the criteria and procedures for addressing risk; the operational structure ensures they are applied consistently throughout the entire service lifecycle. Within each pillar and the operational structure, we have also developed concrete tools for implementation.

The Three Pillars and Operational Structure of NAVER AI Safety Framework 2.0



The first pillar, **Defining and Classifying Risk**, systematically organizes the risks to be addressed in AI services. It identifies users, the public, and the AI service ecosystem as subjects of protection, and combines them with three protected values—protecting life and physical safety, protecting economic value, and preventing unjust discrimination—to form the AI Risk Taxonomy.

The second pillar, **Identifying and Assessing Risk**, examines the level of impact an AI service may have on subjects of protection and protected values. Through the AI Impact Assessment Matrix, safety measures are strengthened based on the expected impact level, taking into account each service's domain and scope.

The third pillar, **Managing and Mitigating Risk**, applies safety measures to identified risks and continuously improves them. Through safety evaluations and user communication, we manage and address risks across the entire AI service lifecycle.

These three pillars are put into practice at the individual service level through our operational structure. The "Consultation on Human-centered AI's Ethics and Safety Considerations 2.0 (CHEC 2.0)" is a company-wide operational structure that spans the entire service lifecycle—from the planning of an AI service through to post-launch evaluation.

The three pillars do not operate in isolation; through the operational structure, they form a connected cycle. The taxonomy provides the basis for assessment; assessment results become the starting point for management and improvement; and newly identified risks during management feed back into the taxonomy. Through this cycle, ASF 2.0 continuously strengthens the safety of AI services in an evolving environment.

3. Defining and Classifying Risk

NAVER is committed to planning and developing AI services that, with safety in mind, do not have a harmful impact on people at any stage.

In particular, to prevent AI—a daily tool meant for humanity—from threatening anyone's life or physical safety, NAVER designs services with safety in mind throughout the entire service lifecycle, conducts evaluations, and continues to oversee safety even after launch.

The safety areas to be addressed in AI services are not entirely new. Rather, they build on the user protection areas that NAVER has long worked to address, with new risk factors unique to AI technology added on top.

Building on the misuse risks defined in "NAVER ASF Beta", we have refined and classified them as risks that may arise within AI services. On this basis, we apply the AI Risk Taxonomy, which designates users, the public, and the AI service ecosystem as subjects of protection and defines a distinct protected value for each.

The Subjects of Protection of ASF 2.0

AI Risk Taxonomy

NAVER organizes the user-centered values articulated in the "NAVER AI Ethics Principles" into three protected values, and continuously strengthens safety through both pre-launch and post-launch impact assessments for AI Safety.

The Protected Values of ASF 2.0

The first is **Protecting Life and Physical Safety**. We will continuously oversee safety to ensure that AI services do not threaten the life or physical safety of users and the public.

The second is **Protecting Economic Value**. We will design and manage AI services to prevent harm arising from the infringement of economic value—whether caused by the AI service itself or by external factors—for users, the public, and the AI service ecosystem.

The third is **Preventing Unjust Discrimination**. We will work to ensure that AI services do not produce unjust discrimination—against any person, including users—that lacks reasonable justification, and we will foster experiences and opportunities where diverse values coexist.

3. Defining and Classifying Risk

The AI Risk Taxonomy Based on the Subjects of Protection and Protected Values of ASF 2.0

Protected Values \ Subjects of Protection	USERS	MEMBERS OF SOCIETY	AI SERVICE ECOSYSTEM
	Protecting Life and Physical Safety	Protecting Life and Physical Safety	Protecting Life and Physical Safety
	Protecting Economic Value	Protecting Economic Value	Protecting Economic Value
	Preventing Unjust Discrimination	Preventing Unjust Discrimination	Preventing Unjust Discrimination

ASF 2.0 categorizes the risks that may arise in AI services by combining subjects of protection with protected values. This AI Risk Taxonomy serves as a starting point for determining which risks to review in the context of how a service is used. Building on this taxonomy, NAVER identifies and assesses risks tailored to the characteristics of each service.

4. Identifying and Assessing Risk

NAVER identifies risks that may arise within a service by systematically organizing both the risks predefined through the AI Risk Taxonomy and those that emerge—or may emerge—during the planning, development, and operation of the service.

The AI Impact Assessment Matrix is a framework for strengthening safety measures based on the level of impact an AI service may have on subjects of protection and protected values.

AI Impact Assessment Matrix

The AI Impact Assessment Matrix considers the entire lifecycle of an AI service. It first distinguishes whether the service falls within a special domain—corresponding to high-risk areas—or a general domain. Services in the general domain are then further divided by whether their scope of use is broad or narrow, and safety measures are determined based on the level of impact on subjects of protection and protected values.

The special domain refers to areas where High-Impact AI, as defined under the AI Basic Act, is used. Safety measures for these services are determined based on the impact on subjects of protection and protected values.

For the general domain, the level of safety measures is determined in consideration of the fact that the impact on subjects of protection and protected values may change as the scope of an AI service expands.

AI Impact Assessment Matrix

		Impact on Subjects and Values of Protection: High	Impact on Subjects and Values of Protection: Low
Special Domain		**High AI Service Risk** Conduct a pre-launch impact assessment, and continue safety evaluations after launch.	**Notable AI Service Risk** In consideration of risks in the special domain, establish safety measures suited to the service context in advance.
General Domain	Broad Scope	**Notable AI Service Risk** In consideration of the risk of use across diverse purposes, establish safety measures suited to the service context in advance.	
General Domain	Narrow Scope		**Low AI Service Risk** After establishing safety measures that take into account the service's characteristics and context, conduct post-launch response and improvement.

For special domains with a high impact on subjects of protection and protected values, we conduct a pre-launch impact assessment and continue safety evaluations after launch.

For special domains with a low impact, or general domains with a high expected impact, we acknowledge the presence of risk and respond by establishing safety measures suited to the characteristics and context of the AI service.

For general domains with a narrow scope of use and a low impact on subjects of protection and protected values, we establish safety measures suited to the service context and then conduct post-launch response and improvement.

5. Managing and Mitigating Risk

NAVER carries out safety evaluations and user communication to manage and address risks across the entire lifecycle of an AI service.

Safety Evaluation

After identifying both the risks predefined through the AI Risk Taxonomy and those that may arise during service operation, we conduct a safety evaluation. This evaluation examines from multiple angles whether the AI service operates as intended and whether it produces harmful or inaccurate outputs.

Based on the results, we determine whether the service meets our safety standards. Even after launch, we conduct ongoing safety evaluations and improvements informed by user feedback and operational insights, enabling us to proactively address emerging risks.

User Communication

NAVER is committed to making AI convenient for everyone. Where AI is used in daily life, we fulfill our accountability to provide users with reasonable explanations. We will continuously improve how and at what level this information is delivered to better serve our users.

When explanations about AI usage are needed, we provide detailed, accessible information tailored to users, and explain in various ways and at different levels how AI is used within the service.

Providing Information on Generative AI When providing explanations and information about AI technology—such as whether generative AI is used—we offer reasonable explanations aligned with user and public awareness, so that users can use the AI service without inconvenience.

Additionally, where generative AI produces outputs that are difficult to distinguish from real content, and there is a risk that users may be misled or harmed, we provide notices in a readily recognizable manner.

Collaboration Through User Feedback NAVER provides channels for users to share feedback or report issues while using AI services. We listen closely to our users and are committed to delivering better services through this process.

6. Operational Structure

The three pillars of ASF 2.0—Defining and Classifying Risk, Identifying and Assessing Risk, and Managing and Mitigating Risk—define the criteria and procedures for addressing risk. The operational structure ensures these are applied consistently across the entire service lifecycle. Through the "Consultation on Human-centered AI's Ethics and Safety Considerations 2.0 (CHEC 2.0)", we systematically track how AI is used across all services and continuously respond to changes in the environment.

Consultation on Human-centered AI's Ethics and Safety Considerations 2.0 (CHEC 2.0)

Published in 2022, the "Consultation on Human-centered AI's Ethical Considerations (CHEC)" was a process for reexamining the human-centered values embedded in AI services through the lens of the "NAVER AI Ethics Principles", and for identifying and addressing concerns about AI.

In establishing ASF 2.0, we strengthened the safety of AI models and services as well as user communication regarding AI-generated content, and developed the "Consultation on Human-centered AI's Ethics and Safety Considerations 2.0 (CHEC 2.0)", which addresses AI Safety alongside AI ethics.

CHEC 2.0 is a company-wide operational structure that provides guidance for individual service teams to implement AI Safety, while enabling collaboration between service teams and the AI Safety Center to meet the level of AI Safety expected by society.

CHEC 2.0 connects the three pillars of ASF 2.0 so that they are concretely implemented in the context of individual services. Through CHEC 2.0, ASF 2.0 functions as a framework in which the results of classification and assessment flow into the operation of each service.

7. AI Safety Governance

NAVER has established a governance structure to put ASF 2.0 into practice. Through collaboration among team members with expertise across diverse fields, we define, identify, assess, and manage the risks of the AI services that NAVER develops and uses.

In particular, in March 2026, we established the AI Safety Center, a dedicated organization for AI Safety within the Chief Corporate Responsibility Officer (CRO) organization of Team NAVER. The AI Safety Center examines and manages AI Safety as one continuous flow, from the model to the service.

ASF 2.0 has advanced the existing AI Safety governance structure into a three-layer management and oversight system. Alongside this, we collaborate with academia, government bodies, and domestic and international AI safety networks, incorporating diverse external perspectives into NAVER ASF.

AI Safety Management and Oversight System

Layer	Responsible Organization	Key Role
Execution (Layer 1)	Service Operations Teams	In the course of planning and operating AI services and developing models, these teams implement ASF 2.0 in accordance with the guidance of the AI Safety Center.
Management and Support (Layer 2)	CRO (Risk Management Working Group) AI Safety Center	The AI Safety Center integrates the perspectives of various departments and is dedicated to AI Safety work across the company. As a dedicated organization, it maintains its independence while supporting AI Safety management and training, so that operational teams can concretely realize AI Safety in the course of putting ASF 2.0 into practice.
Oversight and Final Decision-Making (Layer 3)	Board of Directors (Risk Management Committee)	The Board of Directors is the final decision-making body for the AI Safety matters that ASF 2.0 seeks to manage and oversee.

8. Sharing Our Experience and Insights

NAVER is committed to organizing our experience and know-how in AI Safety into various deliverables and sharing them with users and the broader public.

The criteria and methods for AI Safety must continuously evolve with changes in technology and society, and a wide range of trial and error is inevitable along the way. NAVER, too, has long reflected on what users truly need in terms of AI Safety while planning and developing services, and has put these reflections into concrete practice.

We will work to ensure that the experience accumulated through this process does not remain merely the asset of a single company, but becomes a shared asset for all of society.

Key Sharing and Collaboration Activities

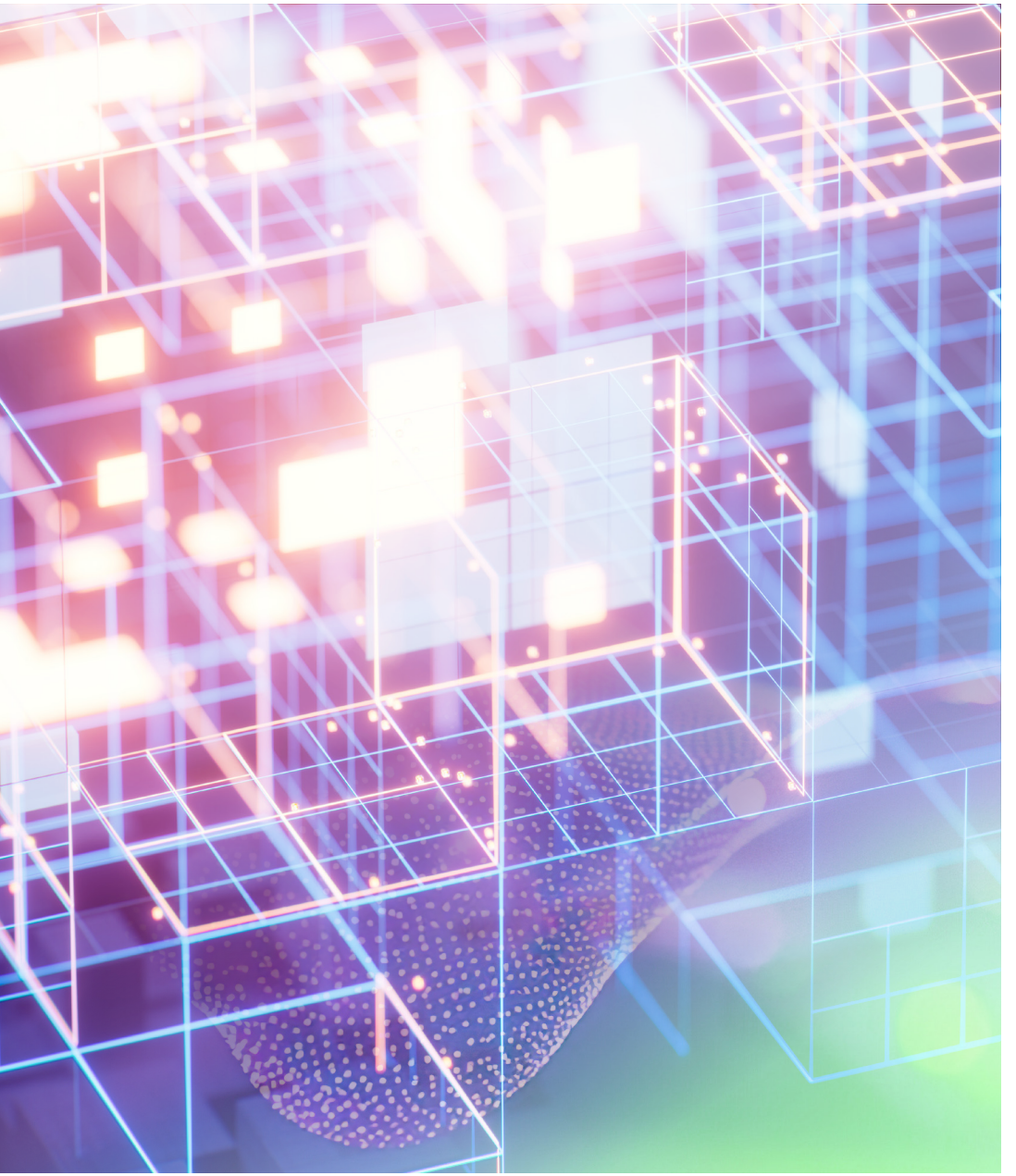
Through diverse collaboration activities, NAVER shares its experience with the public and works with academia, policy circles, and industry to develop standards for AI Safety. We will continue to share our experience and insights so that all users and the broader community can have a safe and trustworthy experience with AI technologies and services.

Sharing AI Safety Activities Through publications such as the NAVER AI Safety Progress Report, we regularly share our AI Safety activities and outcomes, and communicate significant policy developments.

Collaboration with Academia and Policy Circles We collaborate with leading universities on research into AI technology and safety, and engage in AI policy collaboration through initiatives such as the Seoul National University AI Policy Initiative (SAPI).

Consultation with External Experts We enhance the credibility of our governance by seeking advice from independent external experts in AI Safety evaluations and key decision-making processes.

Domestic and International AI Safety Networks We collaborate with domestic institutions such as the AI Safety Institute and the Telecommunications Technology Association of Korea (TTA), as well as international networks such as the UN Global Compact, to participate in global AI Safety discussions and share the experience we have built within Korea's sociotechnical context with the international community.



9. Our Path Forward

The AI that NAVER develops and uses is a daily tool for humanity. We have built technology and services that give connections greater meaning through diversity, opening up a wide range of opportunities and possibilities for users. Going forward, we will continue to prioritize human-centered values in developing and using AI, and will implement NAVER ASF to prevent and mitigate the risks associated with AI services.

AI technology continues to evolve. AI services are moving beyond interacting with users and providing information, expanding into areas where they autonomously reason and act within the scope delegated by users—as in the case of AI agents. At NAVER, we are also introducing agents that help users across a range of areas, including AI Shopping Agent.

As the context in which AI services are used changes, the types and scope of the accompanying risks inevitably change as well. NAVER will respond to these changes by examining risks that need to be newly defined, through the defining and classifying, identifying and assessing, and managing and mitigating of risks.

NAVER recognizes that while AI can make our lives more convenient, it is not perfect—like everything else in the world. We will continuously oversee and improve our AI so that it can serve as a daily tool for humanity.


NAVER