1. AI Safety Framework

As consumer electronics (CE) increasingly integrate on-device AI, AI safety has become an important issue. To address AI safety and related issues, we have developed a comprehensive on-device AI Safety Framework, which focuses on two key processes and four key components

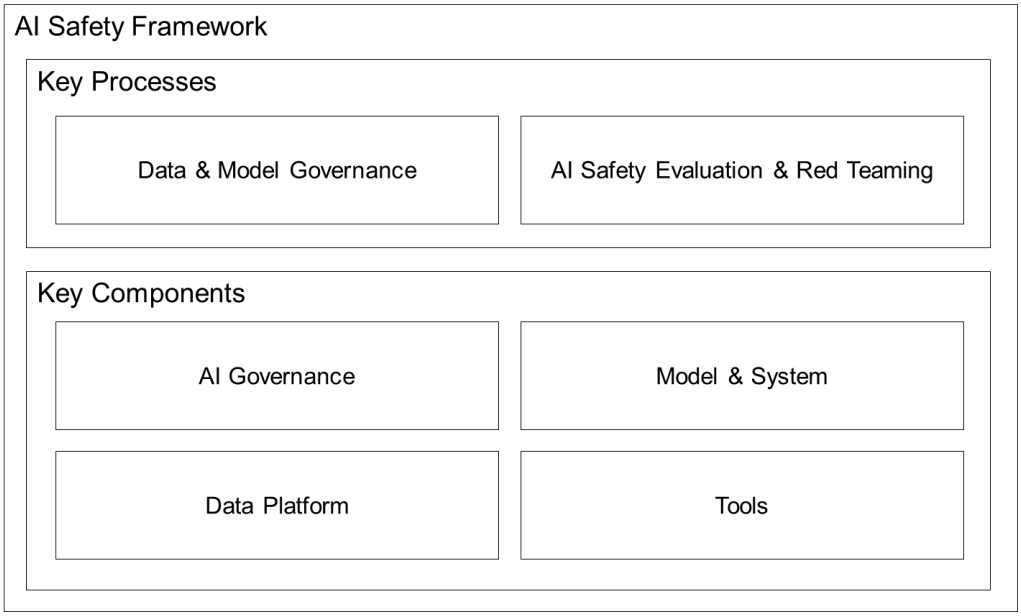

Figure 1 Structure of AI Safety Framework

Importance of AI Safety Framework for CE Devices

With on-device AI, safety is essential for protecting user privacy, ensuring AI's reliability, and mitigating potential risks from malicious attacks. Our AI Safety Framework aims to address these concerns holistically by keeping Fairness, Transparency and Accountability. We aim to ensure safety through a structured framework, which is fundamental to preserve user trust and comply with international standards.

Key Processes to Ensure AI Safety

In the AI Safety Framework, the following processes are implemented:

Data & Model Governance

When acquiring or collecting data, we take steps to evaluate its quality and appropriateness and to address any potential privacy and other issues. The data acquisition process consists of four components: (i) data specification planning and validation, including steps to identify data's characteristics and conduct risk and quality assessment, (ii) closely monitoring data acquisition procedure, (iii) data verification, including steps to keep records of data information and (iv) data registration. After the data is registered, we utilize data cards to manage and keep track of the data.

Further, we also utilize model cards and model evaluation to identify and mitigate risks for a particular AI model. In doing so, we maintain a safety principle checklist and conduct fairness, transparency and accountability assessment.

AI Safety Evaluation & Red Teaming

Each AI model undergoes a safety assessment to identify and mitigate potential risks. This evaluation leverages public safety benchmarks to quantitatively evaluate safety risks in various aspects such as toxicity, bias, truthfulness and overall reliability to prevent unintended outcomes.

We also use controlled adversarial techniques to simulate potential security threats, allowing the company to strengthen model defenses and better prepare for real-world risks.

To address these risks proactively, we have developed an AI Red Teaming process aimed at identifying and addressing safety threats in our AI models.

The AI Safety Evaluation & Red Teaming process builds upon the existing security development life cycle and operates in four phases.

Planning: Through internal review with AI model and system developers, we determine potential AI threat models and plan evaluations accordingly. We focus on setting clear goals, defining scope, identifying potential risks, and establishing resources and strategies necessary for effective testing.

Analysis & Design: We conduct risk assessment of the target AI model, identify the possible risks and extract the requirements for safety evaluation. We design empirical tests which includes making a decision on the appropriate AI safety benchmark datasets and reference AI models based on our risk assessment. In addition, for Red Teaming, we thoroughly analyze security requirements, check the system under test, perform threat modeling, and design hostile tests to identify potential vulnerabilities.

Evaluation & Testing: We conduct evaluation of our target AI model and compare it with reference models. We also set up a separate and isolated Red Team test environment to prevent potential interference with the production system. We execute both manual and automated tests. Interim reports are issued to AI development teams to collect feedback and address any misalignments in the testing goals and results.

Operation & Maintenance: After final validation reports are generated, our evaluation team makes a plan for re-assessment to ensure that the identified vulnerabilities have been addressed promptly and effectively. Before re-assessment, the development team considers possible mitigation plans, such as retraining models or selecting a more safe model to deploy, and adopts the most appropriate one.

This comprehensive approach is intended to safeguard our users and us, fostering a trustworthy environment for deploying AI.

Key Components to Support AI Safety Process

To support key processes, the framework is structured into four key components:

Model & System: We take steps to ensure that AI models are designed and managed to meet safety standards. This includes conducting periodical model evaluation and validation processes throughout AI model's development and lifecycle.

Data Platform: We have implemented measures for data handling and processing, including steps to secure and manage data used in training and deploying AI models.

Tools: We utilize integrated technical resources to manage, monitor, evaluate, and support the safe operation of AI models.

AI Governance: We maintain internal policies and governance for data/model use, control and management.

Governance Operation

We have established an operational team to monitor and manage AI Safety Framework execution. **The AI Strategy Team** manages data, model life cycle, and safety-related processes in connection with various related departments and updates processes in response to applicable laws and regulations.