Technical robustness and safety

The second of the seven requirements of the Ethics Guidelines for Trustworthy AI refers to **Technical Robustness and Safety**.

To prevent harm, AI systems have to be developed with a preventative approach to risks. AI systems need to be **resilient** and **secure**. They need to be safe, ensuring a **fallback plan** in case something goes wrong, as well as being **accurate**, **reliable,** and **reproducible.**

Figure 1 Technical robustness and safety

The assessment of this key requirement is more technical and involves evaluating several critical aspects to ensure that these systems are reliable, secure, and resilient against potential risks. These aspects include:

Reliability and Accuracy

Reliability and Accuracy of AI systems are closely interlinked with technical robustness. They usually involve testing the system's performance across various scenarios, against **different benchmarks, datasets, and use cases**, to ensure it consistently produces accurate and reliable results under different conditions. Metrics such as precision, recall, and F1-score are commonly used to quantify performance.

Evaluating the system's accuracy and reliability helps **in assessing the potential harm due to inaccurate predictions**, including their inherent risk to the system and its users while ensuring the system's behaviour is verified under various conditions.

Security

Security assessment focuses on ensuring that AI systems are protected against **cyber threats and adversarial attacks**, among others. Assessing potential forms of attacks can enhance robustness against attacks designed to deceive or manipulate the AI system. Different security measures or systems should be in place to ensure integrity and resilience against attacks**. Continuous monitoring and updating** of these measures are crucial to adapt to evolving threats.

Figure 2: Data Safety

Resilience

AI systems should be evaluated for their ability to **handle major problems**, such as widespread failures, and how they interact with other systems. By simulating different challenging and stress-testing scenarios, we can identify weak spots and ensure it can recover from significant disruptions.

To make sure AI systems can deal with errors and keep working properly, they should be tested for their **ability to function even when parts of the hardware fail**, **software has bugs, or they receive unexpected inputs**. This includes setting up backup systems to keep things running smoothly if something goes wrong, making the AI system fault-tolerant.

It's also crucial to **develop fallback plans in case problems arise**. Testing the system in various challenging scenarios helps understand and reduce risks, assess potential damage, and develop measures to keep the system safe and operational.

Data Quality and Integrity

There are several dimensions of data quality that are crucial for the robustness of an AI system that uses them. This involves **data accessibility, consistency, completeness, conciseness, reliability, and relevancy, as well as ensuring data privacy and security**. Data validation processes should be in place to maintain high data qualityand thus set a solid foundation for a more robust and safer AI system.

Validation and Verification

Thorough, benchmarking, validation, and verification processes are necessary to ensure that AI systems meet their specifications and requirements, providing also the necessary evidence for assessing and minimizing harm from inaccuracies. This includes **unit testing, integration testing, and system testing**. Independent audits and third-party evaluations can provide additional assurance of the system's robustness and safety.

Monitoring and Maintenance

Continuous monitoring and maintenance are vital for sustaining the robustness and safety of AI systems over time. This involves **continuously tracking the system's performance, detecting anomalies, and addressing issues promptly**. Regular updates and patches should be applied to mitigate new vulnerabilities and improve the system's resilience, whereas a sufficient fallback plan should be available to address major problems not foreseen.

Technical Robustness and Safety in MANOLO

Figure *3:* MANOLO *project*

In MANOLO several partners contribute to ensure AI systems' technical robustness and safety. Under the **guidance of the Z-Inspection® process**, technical partners will collaborate to address the above critical aspects to ensure that these systems are **reliable, secure, and resilient against potential risks.**

In addition, the MANOLO benchmarking framework, led by the **National Centre for Scientific Research "Demokritos"**, will support the validation and verification of the MANOLO AI systems. Dedicated sub-components on model performance (led by **Atos IT Solutions and Services Iberia**) and robustness (led by **FOUR DOT INFINITY INFORMATION AND TELECOMMUNICATIONS SOLUTIONS**) will provide the necessary tools for stress-testing (led by **UPC Universitat Politècnica de Catalunya**)of all envisioned functionalities.

Finally, a set of Data Quality Estimation methods (led by the **National Centre for Scientific Research "Demokritos"**) will ensure that all the data used in the context of MANOLO are of high quality and reliable.

Wrap Up

To summarise, the "Technical Robustness and Safety" key requirement in the Ethics Guidelines for Trustworthy AI **ensures that AI systems are reliable, secure, and resilient against risks**. Assessment for this requirement focuses mainly on technical critical aspects that have to do with reliability and accuracy, security,

resilience, and data quality. Validation and verification processes as well as mechanisms for monitoring and maintenance are also crucial to be present to ensure credible and long-lasting results.

Within the MANOLO project, all technical partners collaborate to **ensure AI systems' robustness and safety, following the Z-Inspection® process** (with the support from **Arcada University of Applied Sciences**) and a dedicated benchmarking framework (led by the **National Centre for Scientific Research "Demokritos"**). This includes tools for **stress-testing and validating AI model performance and robustness.**