Table of Contents

What is Trustworthy AI and why is establishing trust necessary?

The development and perception of many AI applications go through a so-called "hype cycle," as outlined by the market research company Gartner[8], for example. At the beginning, a new technology triggers a trend that can quickly become a hype. While in the hype phase, this technology is often overestimated and potential future scenarios are drafted that cannot yet be realized, a return to reality can be observed after some time. In the case of AI solutions, it is usually only after a completed hype cycle that it becomes clear whether the technology is really innovative, robust enough, and adapted for further sustainable development and market acceptance. This volatility often leads to social uncertainty.

Companies using AI technologies are therefore faced with the challenge of pushing the boundaries of what is possible with the help of the latest technology, while at the same time creating trust for the (AI) services in use. This trust arises, amongst other factors, from ethical frameworks that define the use of AI in the overall social system.

Due to tensions in this area, a political process was initiated at the European level in 2018 that measures artificial intelligence against the backdrop of its "trustworthiness" (Trustworthy AI). As a result, ethical guidelines for Trustworthy AI were established in 2019. They further define the term "trustworthiness" in relation to AI applications in seven different action areas and can therefore be used as an (indirect) definition of the term.

Trustworthy AI occurs when

- AI software components are traceable and controllable by humans
- An environment is established that is protected from unwanted access and interference
- Principles of data protection and data management are followed
- Fundamental rights are respected while promoting diversity, non-discrimination, and fairness
- The focus is on environmental and social well-being
- Sufficient transparency is established, meaning results are verifiable and adaptable while providing the necessary legal remedies to legitimize the use of AI

The importance of setting these ethical and technical guidelines become obvious when looking at the numerous IT, philosophical, and legal challenges that arise in connection with AI applications.

Delineating the AI component

Much of the software used in businesses today control a variety of different processes. Artificial intelligence is an important part of these systems. Here, machine learning models, especially those of deep neural networks, are of integral importance. They are enriched with training data and can then be applied to new, possibly unknown, data of the same type. With the trained models, the AI can then make data-based decisions without any fixed rules established by humans beforehand. The AI component of a software thus influences human-machine interaction more than other parts of the system.

[8] https://www.gartner.com/smarterwithgartner/2-megatrends-dominate-the-gartner-hype-cycle-for-artificial-intelligence-2020/

Establishing Trustworthy AI through audit according to AIC4

The AI Cloud Service Compliance Criteria Catalogue (AIC4)

The AI Cloud Service Compliance Criteria Catalogue (AIC4) is the world's first concrete set of criteria with operationalizable requirements for testing AI applications published by an official governmental institution. It contains AI-specific criteria that allow for an assessment of the trustworthiness of an AI service throughout its lifecycle. The criteria establishes a baseline level of security that can reliably be assessed by independent auditors.

AIC4 was developed specifically for application in the current state of AI technology and applies to cloud-based AI services that rely on machine learning methods and use training data for iterative improvement. Typical application areas for the above methods are speech recognition and language processing services (NLUs & NLPs), image classification tools, (economic) forecasting tools, and scoring models.

In a Conversational AI system, a set of specific conversational functions are provided, which users can then use to model their own virtual assistants (chatbots or voicebots), including desired conversational flows, and configured speech understanding. Central to an AI system for conversational intelligence are the NLU (Natural Language Understanding) functions. These are used to process speech input and evaluate the intentions of that input in multilingual environments. Through platforms such as Cognigy.AI, users are empowered to configure the system to connect the input and output interfaces and train their individual NLU models. The NLU models used in Conversational AI and the data processing in machine learning models make AI applications like Cognigy.AI particularly amenable to auditing by AIC4.

The AIC4 criteria catalogue developed by the German Federal Office for Information Security (BSI) is a response to demands from market participants, the German government and the EU Commission to establish transparency, traceability, and robustness of AI applications. When auditing according to AIC4, companies and providers of AI solutions assume an international pioneering role in the race for Trustworthy AI applications. This results in a competitive advantage in both end-customer deployments and the B2B sector.

Most AI services offer their products - at least optionally - as Software-as-a-Service (SaaS). Cognigy.AI can also be used as a SaaS service. Therefore, the secure operation of AI services in the cloud is an important component of the AIC4 catalogue. With the BSI's C5 criteria catalogue (Cloud Computing Compliance Criteria Catalogue), minimum requirements for secure cloud computing already exist. Therefore, the AIC4 criteria build upon the C5 catalogue and specify it further in the field of the AI lifecycle.

Companies that use AI applications such as Cognigy.AI as an on-premise solution and consequently not as a cloud service can use the AIC4 catalogue to better assess the trustworthiness of the AI services set up on their own or rented IT infrastructure, and can also have them certified by accredited auditors according to the catalogue specifications. In addition to greater transparency and control, competitive advantages and positive image effects can arise for participating companies.