Functional Trustworthiness of AI Systems
Die funktionale Vertrauenswürdigkeit von KI-Systemen

General Info
Allgemeine Information

Functional trustworthiness encompasses all aspects of trustworthiness that are directly dependent on the specific properties of a machine learning system (ML system). Establishing testable quality requirements is crucial for building trust in an AI system's ability to reliably perform its intended functions within a defined application domain. The concept is grounded in the statistically valid testing of AI systems using random samples from the application domain, ensuring that the system's performance meets its own promises and predefined standards. To effectively assess the functional trustworthiness of an AI system we need:

1) the stochastic application domain definition
2) risk-based minimum performance requirements and
3) statistically valid testing

The goal of functional trustworthiness is to build confidence in the functional quality of an AI system. Note that there are other aspects of trustworthiness such as security and oversight that are not covered by functional trustworthiness assessments.

Funktionale Vertrauenswürdigkeit umfasst jene Aspekte der Vertrauenswürdigkeit, die direkt von den spezifischen Eigenschaften des zugrundeliegenden machine-learning-Modells des KI-Systems abhängen. Die Festlegung testbarer Qualitätsanforderungen ist entscheidend, um Vertrauen aufzubauen in die Fähigkeit eines KI-Systems, seine vorgesehenen Funktionen zuverlässig in einem definierten Anwendungsbereich zu erfüllen. Das Konzept basiert auf statistisch validen Tests von KI Systemen anhand zufälliger Stichproben aus dem Anwendungsbereich, um sicherzustellen, dass die Leistung des Systems den behaupteten Spezifikationen und den vordefinierten Standards entspricht. Um die funktionale Vertrauenswürdigkeit eines KI-Systems effektiv zu bewerten, benötigen wir:

1) die stochastische Definition des Anwendungsbereichs,
2) risikobasierte Mindestleistungsanforderungen und
3) statistisch valide Test.

Das Ziel der funktionalen Vertrauenswürdigkeit besteht darin, Vertrauen in die funktionale Qualität eines KI-Systems aufzubauen. Daneben gibt es auch andere Aspekte von Vertrauenswürdigkeit wie beispielsweise Sicherheit und Aufsicht, die nicht Teil von der funktionalen Vertrauenswürdigkeit abgedeckt werden.

Why do we need functional trustworthiness assessments?
Warum brauchen wir funktionale Vertrauenswürdigkeit?

Functional trustworthiness is crucial in data-driven programming because it ensures that AI systems perform reliably. Unlike traditional software written by humans, AI systems are derived from a large amount of complex, high-dimensional data. In machine-learning-based AI systems it is simply not possible to verify their functionality in a formal way, meaning we must rely on the concept of functional trustworthiness. This is the only way to assess the functionality of AI systems.

The concept involves setting clear performance criteria, conducting rigorous statistical testing, and assessing risks to ensure safety and accountability. It provides a framework for verifying that these systems meet its own functional promises and necessary standards, even if they are further developed, thus protecting against unexpected outcomes and ensuring ethical and transparent deployment.

Die funktionale Vertrauenswürdigkeit ist in der datengetriebenen Programmierung von entscheidender Bedeutung, da sie sicherstellt, dass KI Systeme zuverlässig arbeiten. Im Gegensatz zu traditioneller Software, die von Menschen geschrieben wird, basieren KI Systeme auf großen Mengen von komplexen, hochdimensionalen Daten. Im Fall von datengetriebenen KI Systemen ist es schlichtweg nicht möglich, ihre Funktionalität auf formale Weise zu überprüfen, weshalb wir auf das Konzept der funktionalen Vertrauenswürdigkeit angewiesen sind. Nur so lässt sich die Funktionalität von KI Systemen bewerten.

Das Konzept umfasst die Festlegung klarer Leistungskriterien, die Durchführung rigoroser statistischer Tests und die Bewertung von Risiken, um Sicherheit und Verantwortlichkeit zu gewährleisten. Es bietet einen Rahmen, um zu überprüfen, ob diese Systeme die versprochenen Funktionen und die erforderlichen Standards erfüllen, selbst wenn diese weiterentwickelt werden, und schützt so vor unerwarteten Ergebnissen und gewährleistet eine ethische und transparente Implementierung.

MORE INFO
MEHR INFOS