Thetis

Ihr Weg zur KI-Sicherheit.

Starter

Preis auf Anfrage

Unsere Basisvariante mit zubuchbaren Optionen. So individuell wie Ihr System.

Thetis Performance	✓
Thetis Fairness	monatl. zubuchbar
Thetis Data Quality	monatl. zubuchbar
Thetis Robustness	monatl. zubuchbar
Thetis Uncertainty	monatl. zubuchbar

Premium

Preis auf Anfrage

Die komfortable Komplettlösung.

Thetis Performance	✓
Thetis Fairness	✓
Thetis Data Quality	✓
Thetis Robustness	✓
Thetis Uncertainty	✓

Premium Self-hosted

Preis auf Anfrage

Auf Wunsch richten wir Ihnen Ihr privates Thetis in Ihrem System ein. Wir unterstützen Sie hierbei gerne bei der Einrichtung, Inbetriebnahme und Wartung. Sprechen Sie uns an!

Thetis GenAI

Preis auf Anfrage

Die technische Bewertung der KI-Sicherheit von generativen KI-Systemen bspw. zur Sprach- oder Bildgenerierung erfordert eine individuelle technische Überprüfung. Buchen Sie jetzt einen Termin mit uns.

Thetis SafeAI Consulting

Preis auf Anfrage

Haben Sie weitere Fragen zu dem Thema der KI-Sicherheit? Benötigen Sie technische Unterstützung bei der Bewertung oder Verbesserung Ihres Systems? Unsere KI-Experten unterstützen Sie gerne.

Stand: 08.08.2024

Es gelten unsere Allgemeinen Geschäftsbedingungen.

e:fs TechHub GmbH

Thetis Performance

Art. 11 und 15 KI-VO
Annex IV Abs. 2 lit. g KI-VO

Klassifikation

Nach Art. 15 ist die Genauigkeit eines KI-Systems vor dem Inverkehrbringen zu bestimmen und in der technischen Dokumentation nach Art. 11 (Technische Dokumentation) Annex IV Abs. 2 lit. g KI-VO zu dokumentieren.

Im Kontext der **Klassifikation** (bspw. anhand von Merkmalen, Texten oder Bildern) berechnet Thetis verschiedene Metriken, die zur Bewertung der Leistungsfähigkeit eines KI-Algorithmus verwendet werden:

- Genauigkeit (Accuracy) ist die Wahrscheinlichkeit, dass das KI-System eine Instanz korrekt klassifiziert.
- Sensitivität (Recall) gibt Aufschluss darüber, inwiefern ein Klassifikator in der Lage ist, Instanzen für verschiedene Labels korrekt zu identifizieren.
- Die Präzision (Precision) gibt an, inwiefern der Klassifikator in der Lage ist, genaue Vorhersagen der jeweiligen Klassen für verschiedene Labels zu treffen.
- Der Matthews Korrelationskoeffizient (MCC) ist eine Metrik zur Bewertung der Leistung eines Klassifikators, bei welcher verschiedene Aspekte in einer einzigen Punktzahl zusammengefasst werden. Die Metrik reicht von -1 bis +1, wobei ein Wert von +1 für perfekte Schätzungen ohne Fehler steht, ein Wert von 0 auf zufällige Klassifikationsentscheidungen hindeutet und -1 eine völlige Unstimmigkeit zwischen Vorhersagen und tatsächlichen Bezeichnungen anzeigt.

Zusammen betrachtet ermöglichen diese Metriken eine umfassende Analyse und Bewertung der Leistungsfähigkeit von Klassifikationssystemen. Die zu erzielende Leistungsfähigkeit ist allerdings von dem jeweiligen Anwendungsfall abhängig, sodass eine für jeden Anwendungsfall eine individuelle Bewertung erfolgen muss (nicht Teil des Pakets).

Thetis Performance

Art. 11 und 15 KI-VO
Annex IV Abs. 2 lit. g KI-VO

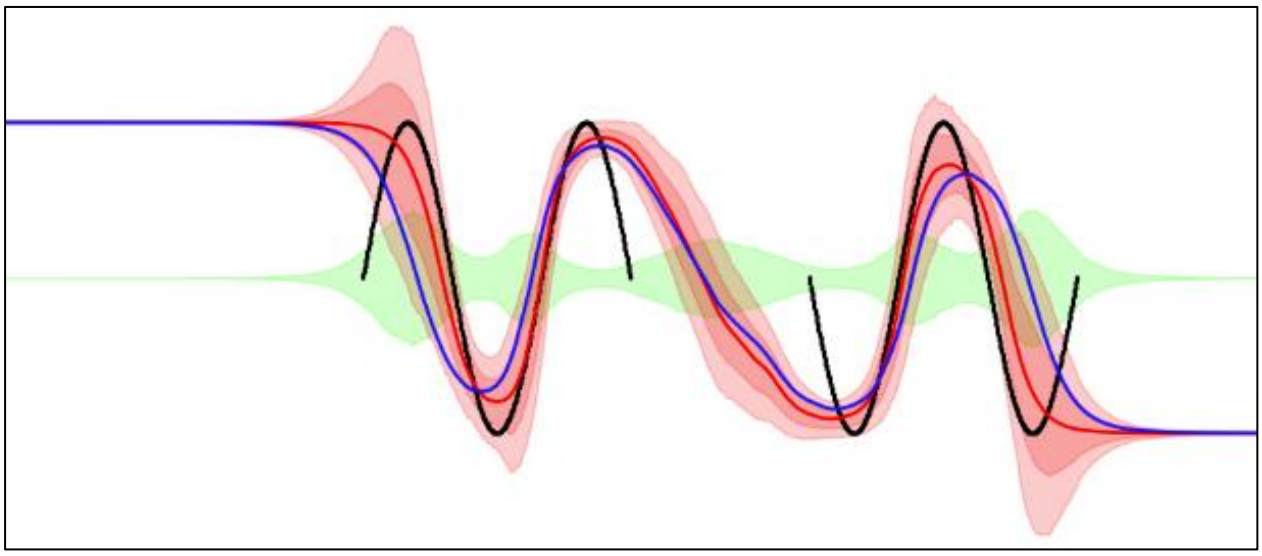
Regression

Nach Art. 15 ist die Genauigkeit eines KI-Systems vor dem Inverkehrbringen zu bestimmen und in der technischen Dokumentation nach Art. 11 (Technische Dokumentation) Annex IV Abs. 2 lit. g KI-VO zu dokumentieren.

Für Anwendungen der **Regression** (Schätzung eines kontinuierlichen Zielwerts, bspw. eines Scores) berechnet Thetis verschiedene Metriken zur Bewertung des KI-Systems:

- Mittlerer absoluter Fehler (Mean Absolute Error, MAE).
- Median absoluter Fehler (Median Absolute Error), robust gegenüber Ausreißern.
- Mittlerer quadratischer Fehler (Mean Squared Error, MSE).
- Wurzel des mittleren quadratischen Fehlers (Root Mean Squared Error, RMSE).
- Mittlerer absoluter prozentualer Fehler (Mean Absolute Percentage Error, MAPE): Mittelwert des absoluten Fehlers relativ zum absoluten Zielwert.
- Bestimmtheitsmaß (R^2): Fehler relativ zur Varianz der Zielwerte.

Eine Bewertung der Metriken und somit der Leistungsfähigkeit ist abhängig von dem Anwendungsfall. Es bedarf somit einer individuellen Analyse, Interpretation und Bewertung (nicht Teil des Pakets).



Thetis Performance

Art. 11 und 15 KI-VO
Annex IV Abs. 2 lit. g KI-VO

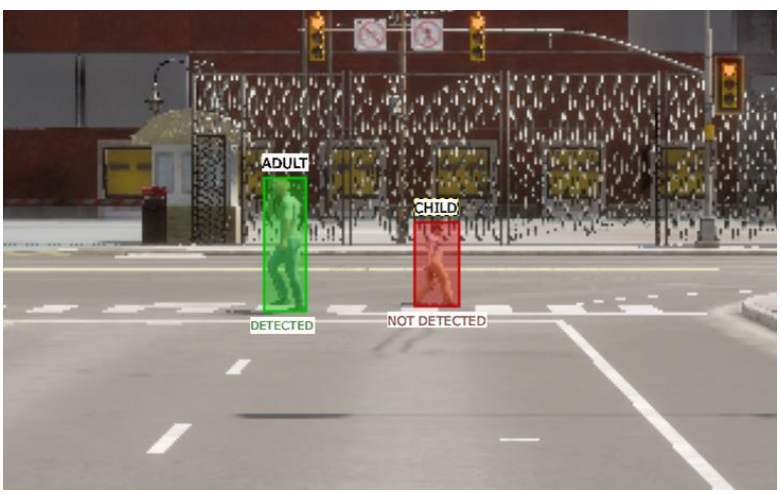
Objektdetektion

Nach Art. 15 ist die Genauigkeit eines KI-Systems vor dem Inverkehrbringen zu bestimmen und in der technischen Dokumentation nach Art. 11 (Technische Dokumentation) Annex IV Abs. 2 lit. g KI-VO zu dokumentieren.

Für KI-Systeme der **Objekterkennung** berechnet Thetis folgende Metriken, die als Grundlage zur Bewertung der Leistungsfähigkeit eines Detektionssystems herangezogen werden können:

- Präzision: Anteil der korrekt vorhergesagten Objekte an allen vorhergesagten Objekten. Diese Metrik misst, wie zuverlässig die getroffenen Detektionen sind, ob sie also realen Objekten entsprechen.
- Recall: Anteil der korrekt identifizierten Objekte an allen tatsächlich existierenden Objekten. Der Recall misst somit die Eigenschaft des Detektionssystems, alle realen Objekte korrekt erkennen zu können.
- F1-Score: Harmonisches Mittel von Präzision und Recall.
- Mean Average Precision (mAP): Durchschnittliche Präzision über alle realen Objekte hinweg. Die mAP misst die Fähigkeit des Systems, Objekte korrekt zu erkennen und ihre Positionen genau zu bestimmen. Im Gegensatz zu reiner Präzision und Recall, die nur einzelne Aspekte der Vorhersagegenauigkeit betrachten, integriert mAP beide Aspekte und liefert somit eine umfassendere Bewertung der Gesamtleistung des Modells.

Eine Interpretation und Bewertung der Metriken muss für jeden Anwendungsfall individuell erfolgen (nicht Teil des Pakets).



Thetis Fairness

Art- 11 und 10 Abs. 2 lit. f KI-VO
Annex IV Abs. 3 KI-VO

Die KI-Verordnung schreibt eine umfassende Untersuchung des Datensatzes auf einen möglichen Bias sowie eine umfassende Untersuchung des KI-Systems auf mögliche Diskriminierung vor. Allerdings lässt die KI-Verordnung die konkrete Ausgestaltung einer konkreten Definition jedoch offen, wie dies technisch konkret analysiert und bewertet werden soll. In der Literatur finden sich hierzu verschiedene konkrete Definitionen, wobei wir für unsere technischen Analysen die Definition der Group Fairness verwenden. Diese Definition schreibt eine Gleichbehandlung von Personengruppen vor, um strukturelle Diskriminierung zu vermeiden und die statistische Unabhängigkeit eines KI-Systems bzgl. sensitiver Merkmale wie bspw. Geschlecht oder Herkunft zu analysieren. Im Rahmen der technischen Analysen wird untersucht, ob ein KI-System für unterschiedliche Personengruppen eine unterschiedliche Leistungsfähigkeit aufweist.

Bei „klassischen“ KI-Systemen (keine GenAI Systeme) kann anhand der Zielwerte („Grundwahrheit“ oder „Ground Truth“), die in dem Evaluationsdatensatz vorhanden sind, die Leistungsfähigkeit des KI-Systems bestimmt werden. Durch die Betrachtung der Leistungsfähigkeiten gruppiert nach unterschiedlichen Personengruppen können somit Aussagen über die Fairness eines KI-Systems getroffen werden.

Voraussetzung für die Bias- und Fairness-Analyse des Datensatzes bzw. des KI-Systems: Annotationen zu den zu untersuchenden geschützten Merkmalen, die keinen Einfluss auf das Ergebnis haben dürfen. Dies ist in Abhängigkeit von dem Anwendungskontext zu evaluieren. Zu dem Zweck der „Erkennung und Korrektur von Verzerrungen im Zusammenhang mit Hochrisiko-KI-Systemen“ ist nach Art. 10 Abs. 5 KI-VO eine Ausnahmeregelung des Art. 9 Abs. 1 DSGVO vorgesehen, dass besonders geschützte Merkmale unter Beachtung zusätzlicher Sicherheitsvorkehrungen verarbeitet werden dürfen

Thetis

Ihr Weg zur KI-Sicherheit.

Thetis Data Quality

Art. 10 Abs. 2 lit. f KI-VO

Zur Bewertung des Bias eines Datensatzes für das Training und die Bewertung „klassischer“ KI-Systeme (keine GenAI Systeme) wird vorausgesetzt, dass die jeweiligen geschützten Merkmale (bspw. Geschlecht oder Herkunft) unabhängig von dem Zielwert sind (statistische Unabhängigkeit). Die geschützten Merkmale, die keinen Einfluss auf den Zielwert haben dürfen, sind je nach Anwendungsfall zu spezifizieren.

Des Weiteren richtet sich die Bewertung eines Datensatzes, welcher zum Training oder zur Evaluation eines KI-Systems verwendet wird, zum einen nach dessen Größe. Ein Datensatz muss eine gewisse Mindestgröße aufweisen, um für das Training eines KI-Systems geeignet zu sein oder um statistisch signifikante Aussagen über die Leistungsfähigkeit eines KI-Systems bei dessen Bewertung zu ermöglichen.

Zum anderen hat die Verteilung der Zielwerte innerhalb eines Datensatzes einen signifikanten Einfluss auf das Training bzw. auf die Evaluation des KI-Systems. Ein potenzieller Bias in dem Datensatz wirkt sich unmittelbar auf das Training und/oder die Evaluation des Systems aus. Idealerweise folgen die Zielwerte innerhalb eines Datensatzes einer Gleichverteilung. Dies wird als ein Kriterium für die Bewertung des Datensatzes mit herangezogen. Dieses Kriterium gilt nicht nur für die Verteilung der Zielwerte, sondern auch für die Verteilung von potenziell vorhandenen geschützten Merkmalen. Auch hier wird ein Datensatz tendenziell schlechter bewertet, sofern eine Personengruppe über- bzw. unterrepräsentiert vorhanden ist.

Weitere Kriterien zur Evaluation eines Datensatzes nach Art. 10 (Daten und Daten-Governance) Abs. 3 KI-VO sind die Relevanz, Repräsentanz, Fehlerfreiheit sowie Vollständigkeit der Daten. Auf Wunsch können wir anwendungsspezifisch geeignete Analysewerkzeuge einsetzen, um die Anforderungen aus der KI-Verordnung individuell erfüllen zu können (nicht Teil des Pakets „Thetis Data Quality“).

Stand: 08.08.2024

e:fs TechHub GmbH

Thetis

Ihr Weg zur KI-Sicherheit.

Thetis Robustness

Art. 15 KI-VO

Der Begriff der Robustheit eines KI-Systems definiert sich aus verschiedenen Teilaspekten der Safety (Betriebssicherheit, Schutz von Personen und Umwelt vor dem System) und Security (Informationssicherheit, Schutz des Systems vor gezielten Angriffen) zusammen.

Unter die Betriebssicherheit fällt die Sicherstellung der Leistungsfähigkeit bei externen Störfaktoren. Bei bildbasierten Verarbeitungsketten können dies bspw. Regentropfen auf der Kameralinse, Helligkeitsänderungen oder Sensorrauschen sein. Für bildbasierte Anwendungsfälle können mit Thetis die folgenden Szenarien automatisiert analysiert und bewertet werden:

- Störanfälligkeit bei Änderungen des Kamerabildes: Helligkeit, Kontrast, Sättigung.
- Störanfälligkeit gegenüber Rauschen: Gauß-Noise, Shot-Noise, Impuls-Noise.
- Störanfälligkeit gegenüber Umwelteinflüssen: Regentropfen auf dem Kamerabild.

Weitere Aspekte der Robustheit umfassen die Informationssicherheit des Systems gegenüber gezielten Manipulationsversuchen von Extern. Als Beispiele nennt Art. 15 Abs. 4 die Manipulation des Trainingsdatensatzes oder die Manipulation des KI-Systems („Adversarial Attacks“). Der Schutz des Systems kann einerseits durch geeignete Maßnahmen aus der Cybersecurity sichergestellt werden. Andererseits kann eine Bewertung des KI-Systems bzgl. der Anfälligkeit für Adversarial Attacks ermittelt werden, um die Risikowahrscheinlichkeit für ein solches Szenario realistisch einschätzen zu können. Die Prüfung eines Systems auf Anfälligkeit für Adversarial Attacks muss jedoch stets individuell angepasst an das jeweilige System erfolgen, da die Prüfverfahren teilweise intrinsisch das KI-System analysieren und somit abhängig von der jeweiligen Architektur sind. Bei Zubuchung von Consulting-Leistungen können wir solche Prüfungen zusammen mit der Kundin bzw. dem Kunden durchführen.

Stand: 08.08.2024

e:fs TechHub GmbH

Thetis

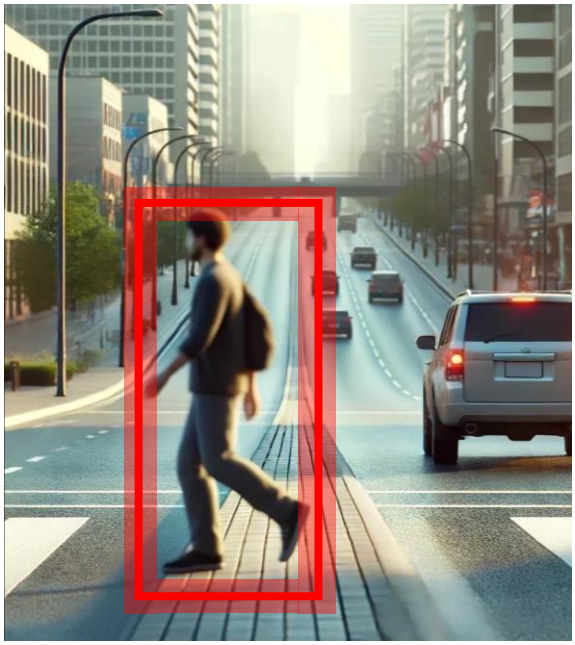
Ihr Weg zur KI-Sicherheit.

Thetis Uncertainty

Art. 15 KI-VO

Der Begriff der Robustheit eines KI-Systems definiert sich aus verschiedenen Teilaspekten der Safety (Betriebssicherheit, Schutz von Personen und Umwelt vor dem System) und Security (Informationssicherheit, Schutz des Systems vor gezielten Angriffen) zusammen.

Ein wesentlicher Aspekt der Betriebssicherheit ist die Zuverlässigkeit der Unsicherheitsschätzung eines KI-Systems. Gängige KI-Systeme prädizieren neben dem gewünschten Ergebnis in der Regel ein Konfidenzmaß, welches angibt, mit welcher Sicherheit das System die Korrektheit seiner Ausgabe selbst einschätzt. Nach der DIN SPEC 92005:2024 ist eine zuverlässige Unsicherheitsschätzung essenziell, um die Betriebssicherheit zu gewährleisten und im Zweifelsfall geeignete Sicherheitsmaßnahmen einleiten zu können, sofern die Konfidenz ein gewisses Maß unterschreitet.



In der Vergangenheit haben zahlreiche Forschungsarbeiten jedoch ergeben, dass dieses Konfidenzmaß von modernen KI-Systemen in der Regel zu hoch eingeschätzt wird und von dem tatsächlich beobachteten Fehler abweicht. Ein Beispiel: ein Objektdetektionssystem erkennt 100 Personen mit einer Konfidenz von je 90%. Es wird somit erwartet, dass 90 von 100 detektierten Personen korrekt erkannt wurden. Bei einer Abweichung der Fehlerrate von der Konfidenz wird von einer Fehlkalibrierung der Unsicherheitsschätzung gesprochen. Thetis ist sowohl in der Lage, relevante Metriken zu berechnen als auch eine automatisierte Bewertung und Einordnung der Ergebnisse vorzunehmen.

Somit sind wir in der Lage, eine aussagekräftige Bewertung der Qualität der Unsicherheitsschätzung vorzunehmen. Eine Voraussetzung ist das Vorhandensein entsprechender Konfidenzinformationen zu den jeweiligen Schätzungen des KI-Systems.

Thetis GenAI

Diskriminative vs. generative KI-Systeme

Die technische Bewertung der KI-Sicherheit von generativen KI-Systemen bspw. zur Sprach- oder Bildgenerierung erfordert eine individuelle technische Überprüfung. Wir unterstützen Sie hierbei gerne. Um zu verstehen warum, erklären wir an dieser Stelle die Unterschiede zwischen „klassischer“ diskriminativen und generativen KI-Systemen.

Diskriminative Systeme können auch als entscheidungsbasierte Systeme aufgefasst werden, welche zur Beantwortung verschiedener Fragestellungen genutzt werden können (bspw. zur Klassifikation von Bildern oder Entitäten, zur Objekterkennung im Straßenverkehr oder zur Schätzung bestimmter Werte, bspw. einer Prognose des Verkehrsaufkommens). Beispiele für diskriminative Modelle sind logistische Regression, Support Vector Machines (SVMs) und neuronale Netze, die zur Klassifikation (Bilder, Datenpunkte, Text), Detektion oder Regression eingesetzt werden.

Generative KI-Systeme hingegen sind darauf ausgelegt, die Datenverteilung selbst zu lernen, um neue, ähnliche Datenpunkte zu erzeugen. Sie beantworten Fragen wie: "Wie sieht ein neues, plausibles Beispiel aus, das aus denselben Daten stammt?" Diese Modelle lernen die Verteilung der Daten, aus denen sie neue Datenpunkte generieren können, und sind in der Lage, neue Beispiele zu erstellen, die den Trainingsdaten ähneln. Beispiele für Architekturen generativer Modelle sind Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) oder Stable Diffusion. Diese Modelle werden verwendet, um Bilder, Texte oder Musik zu erzeugen.

Die Einteilung eines KI-Systems in eine dieser beiden Kategorien hat unmittelbare Auswirkungen darauf, wie dieses System bewertet werden kann. Die Leistungsfähigkeit diskriminativer KI-Systeme wird typischerweise ermittelt, indem die Prädiktionen eines Systems auf einem vorhandenen Evaluationsdatensatz mit den jeweiligen „wahren“ Zielwerten verglichen werden (Beispiel: Vergleich der prädizierten Klasse im Vergleich zu der bekannten Klasse bei der Bildklassifikation). Es gibt bei der Auswertung diskriminativer KI-Systeme somit ein eindeutiges Evaluationsziel.

Die Auswertung generativer KI-Systeme gestaltet sich jedoch wesentlich komplexer, da die Ausgabe häufig Interpretationsspielraum zulässt und es kein „eindeutiges“ Ergebnis gibt. Somit ist die Auswertung der Leistungsfähigkeit generativer KI-Systeme zumeist an den Anwendungskontext gebunden und erfordert erheblichen Mehraufwand.