

- openHPI: ChatGPT: Was bedeutet generative KI für unsere Gesellschaft? -

# Erklärbarkeit und KI Sicherheit

Johannes Hötter  
Christian Warmuth

# Begriff der Erklärbarkeit (engl. Explainability)

Erklärbarkeit in der KI bezieht sich auf die Fähigkeit, die internen Mechanismen oder Entscheidungen eines KI Systems nachzuvollziehen.



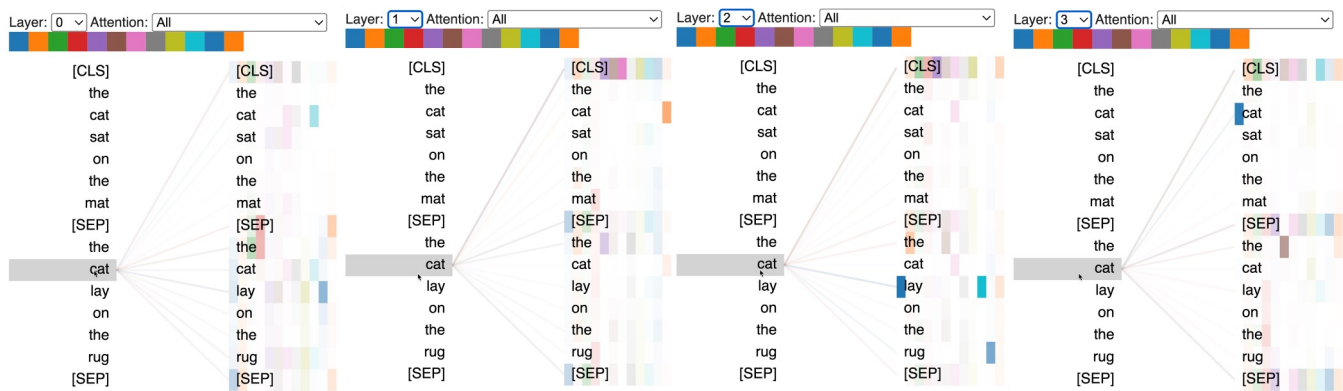
## Erklärbarkeit und KI Sicherheit

Johannes Hötter,  
Christian Warmuth

openHPI

# Intrinsic Explainability

## Attention-Based Intrinsic Explainability



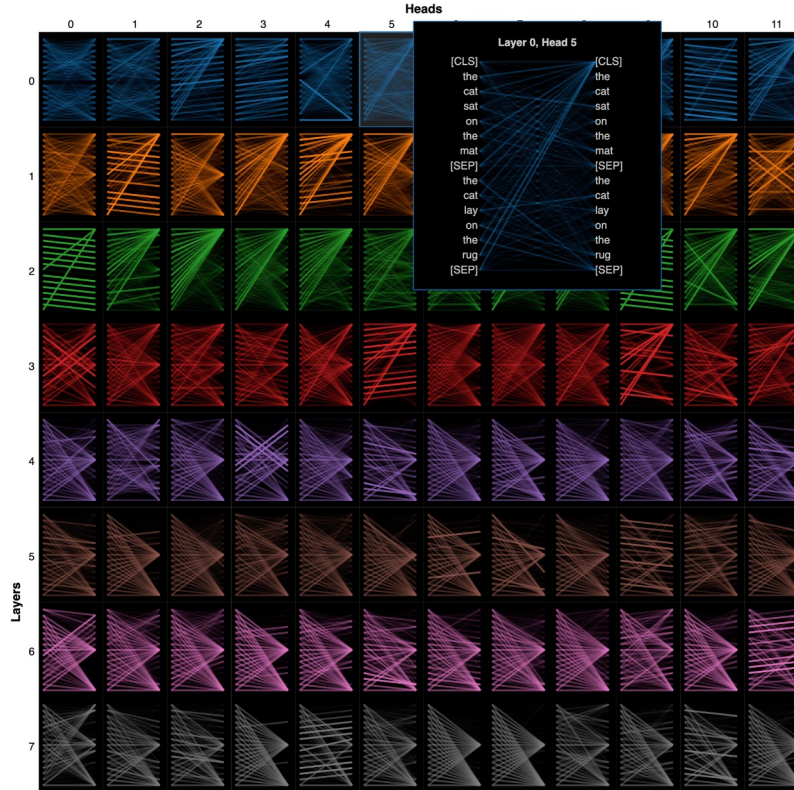
Quelle: <https://github.com/jessevig/bertviz>

**Erklärbarkeit und  
KI Sicherheit**

Johannes Hötter,  
Christian Warmuth

openHPI

# Intrinsic Explainability



Quelle: <https://github.com/jessevig/bertviz>

**Erklärbarkeit und  
KI Sicherheit**

Johannes Hötter,  
Christian Warmuth

openHPI

# Intrinsic Explainability

## Saliency-based Intrinsic Explainability

Beispiel: Wörter werden entfernt oder hinzugefügt und die Änderungen werden beobachtet.



Quelle: Aleph Alpha



# Intrinsic Explainability

## Saliency-based Intrinsic Explainability

Beispiel: Wörter werden entfernt oder hinzugefügt und die Änderungen werden beobachtet.

New Delhi is the capital of >> India

**Erklärbarkeit und  
KI Sicherheit**

Johannes Hötter,  
Christian Warmuth

openHPI

# Intrinsic Explainability

The screenshot shows the Aleph Alpha Playground interface. At the top, there are navigation links: Profile, Playground, Support, and Admin. The main header includes a logo, a sun icon, and buttons for Sign out and Buy Credits. The Playground section has tabs for Complete, Explain (active), Q & A, and Summarize. On the left, the Settings panel includes: Language (Spanish and Italian), Prompt Granularity (Sentence), Target Granularity (Token), Postprocessing (None), a checked Normalize checkbox, Control Factor (1.4), Control Token Overlap (Partial), and checkboxes for Control Log Additive and Use Contextual Control Threshold. The main area displays a text prompt with several segments highlighted in orange. Below the prompt is a Target field and an Edit button.

Profile ▾ Playground ▾ Support ▾ Admin

⦿ ☀ Sign out Buy Credits

**Playground** Complete Explain Q & A Summarize

**Settings**

Spanish and Italian

Prompt Granularity ⓘ  
Sentence

Target Granularity ⓘ  
Token

Postprocessing ⓘ  
None

☒ Normalize ⓘ

Control Factor ⓘ  
1.4

Control Token Overlap ⓘ  
Partial

☐ Control Log Additive ⓘ

☐ Use Contextual Control Threshold

Prompt

Negative Positive

the attention of gentlemen like Mr. Ricci and his colleagues, despite the almost certain fact that he hides a fortune of indefinite magnitude somewhere about his musty and venerable abode. He is, in truth, a very strange person, believed to have been a captain of East India clipper ships in his day; so old that no one can remember when he was young, and so taciturn that few know his real name. Among the gnarled trees in the front yard of his aged and neglected place he maintains a strange collection of large stones, oddly grouped and painted so that they resemble the idols in some obscure Eastern temple. This collection frightens away most of the small boys who love to taunt the Terrible Old Man about his long white hair and beard, or to break the small-paned windows of his dwelling with wicked missiles; but there are other things which frighten the older and more curious folk who sometimes steal up to the house to peer in through the dusty panes. These folk say that on a table in a bare room on the ground floor are many peculiar bottles, in each a small piece of lead suspended pendulum-wise from a string. And they say that the Terrible Old Man talks to these bottles, addressing them by such names as Jack, Scar-Face, Long Tom, Spanish Joe, Peters, and Mate Ellis, and that whenever he speaks to a bottle the little lead pendulum within makes certain definite vibrations as if in answer. Those who have watched the tall, lean, Terrible Old Man in these peculiar conversations, do not watch him again. But Angelo Ricci and Joe Czaneek and Manuel Silva were not of Kingsport blood; they were of that new and heterogeneous alien stock which lies outside the charmed circle of New England life and traditions, and they saw in the Terrible Old Man merely a tottering, almost helpless greybeard, who could not walk without the aid of his knotted cane, and whose thin, weak hands shook pitifully. They were really quite sorry in

Target

The old man is described as exceedingly feeble both physically and **mentally**.

Edit

Quelle: Aleph Alpha

**Erklärbarkeit und  
KI Sicherheit**

Johannes Hötter,  
Christian Warmuth

openHPI

# Extrinsic Explainability

Möglich wenn ein LLM Zugang zu einer externen Wissensbasis hat  
(z.B. VectorStore mit Fakten)

Welche Städte in Amerika waren Gastgeber  
der Olympischen Spiele?

Lake Placid und Los Angeles.

**Seite 5: Geschichte der  
Olympischen Spiele.**

Zwei Städte in Amerika  
haben bereits die  
Olympischen Spiele...

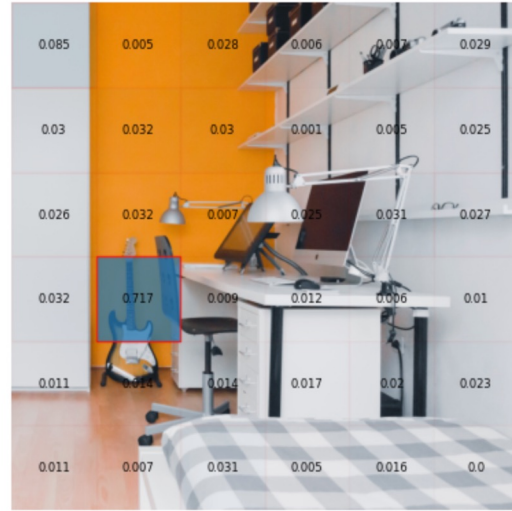
**Erklärbarkeit und  
KI Sicherheit**

Johannes Hötter,  
Christian Warmuth

openHPI



# Multimodale Erklärbarkeit



<b>The</b> 0.077	<b>instrument</b> 1.0	<b>on</b> 0.112	<b>the</b> 0.051	<b>picture</b> 0.117	<b>is</b> 0.001	<b>a</b> 0.001	guitar.
---------------------	--------------------------	--------------------	---------------------	-------------------------	--------------------	-------------------	---------

## Erklärbarkeit und KI Sicherheit

Johannes Hötter,  
Christian Warmuth

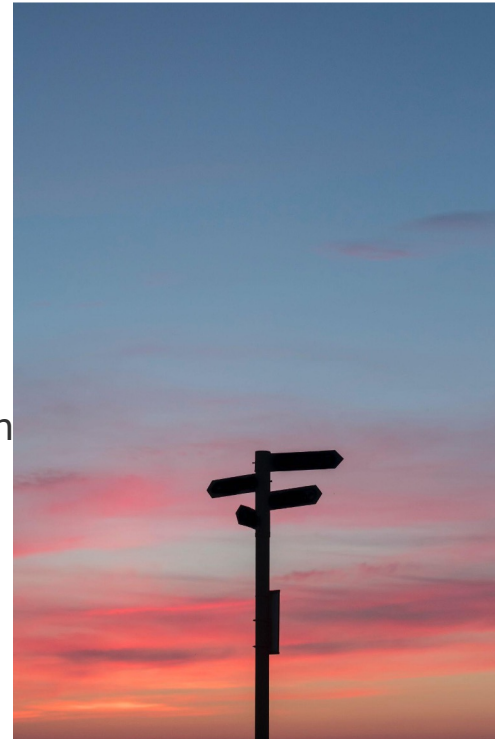
openHPI

Quelle: Aleph Alpha

# Begriff der KI Sicherheit

KI Sicherheit ist ein multi-disziplinäres Feld. Wir haben bereits einige Aspekte der KI Sicherheit beleuchtet:

- Fairness
- Transparenz und Erklärbarkeit
- Schutz vor "Angriffen"/Ungewolltem Verhalten
- ...



## **Erklärbarkeit und KI Sicherheit**

Johannes Hötter,  
Christian Warmuth

openHPI

# Besondere Wichtigkeit

KI Systeme werden zunehmend von isoliert agierenden Systemen zu aktiv agierenden Systemen (teilweise autonom agierenden Systemen).

Beispiele: Zugriff auf das Internet nicht nur in lesendem Modus, API-Calls,...  
(Siehe Beispiel in Woche 2: Agenten)

Umso wichtiger sind in diesem Kontext alle Anstrengungen, KI Sicherheit zu gewährleisten.



## Erklärbarkeit und KI Sicherheit

Johannes Hötter,  
Christian Warmuth

openHPI

- openHPI: ChatGPT: Was bedeutet generative KI für unsere Gesellschaft? -

# Erklärbarkeit und KI Sicherheit

Johannes Hötter  
Christian Warmuth