



VERTRAUEN IST GUT - KONTROLLE IST BESSER!

KI-Systeme, die maschinelles Lernen nutzen, sind häufig „Blackboxes“. Da sie auf großen Datenmengen trainiert wurden, kann man oft nicht genau sagen, wie ein solches KI-System zu einem Ergebnis gekommen ist. Das ist ein Problem, denn wie können die Ergebnisse von KI-Systemen dann kontrolliert werden? Natürlich mit anderen KI-Systemen! Diese nutzen Methoden der sogenannten **erklärbaren Künstlichen Intelligenz (explainable Artificial Intelligence, kurz XAI)**.



Das Problem:

Bei normalen Algorithmen können wir alle Anweisungen nachvollziehen und wissen so ganz genau, wie Programme etwas berechnen. Bei KI-Algorithmen wie beispielsweise Neuronalen Netzen (mehr dazu erfährst Du in Station 03) ist das nicht mehr so einfach: Das Netz führt eine komplexe Berechnung mit sehr vielen Zahlen durch, die alle in das Netz eingegebenen Informationen darstellen. Daraus entsteht eine Ausgabe.

Dabei ist allerdings **für Menschen nicht mehr klar, welche Zahl welche Information darstellt und wie wichtig sie ist**. Deshalb können sogar die Entwickler des KI-Systems **nicht überprüfen, ob ein KI-System die richtigen Muster und Regeln in einem Datensatz gefunden hat** oder ob etwas Falsches gelernt wurde. Es kann passieren, dass das KI-System zwar gut auf den Trainingsdaten, also auf den Beispielen, die es zum Lernen genutzt hat, funktioniert, aber nicht mehr für neue Eingaben. Das nennt man auch **Overfitting (Überanpassung an die Trainingsdaten)**.

Ein anderes Problem ist, dass die **Trainingsdaten nicht ausgewogen** sind, sondern für bestimmte Personengruppen nur sehr spezielle Daten vorhanden sind. Zum Beispiel zeigt ein Datensatz aus Bildern von Jugendlichen hellhäutige Teenager häufig beim Sport, während dunkelhäutige häufig bei Verhaftungen gezeigt werden. Dieses Problem wird **Sampling Bias (Verzerrung bei der Auswahl der Trainingsdaten)** oder **unfairer Bias** genannt.



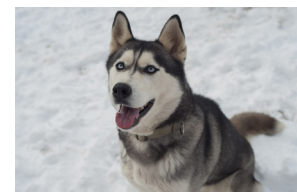
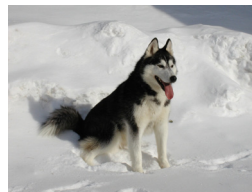


Was bedeutet das jetzt?

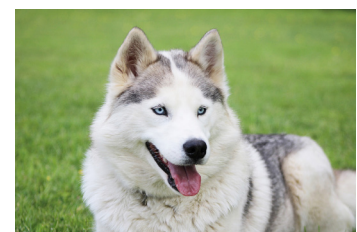
Wir arbeiten mit **zwei KI-Systemen**. **KI-System 1** nutzt maschinelles Lernen, um aus Daten ein Modell zu erstellen. Dieses Modell ist eine Blackbox, man kann also nicht genau feststellen, wie es zu seinem Ergebnis kommt.

Wir wollen ein Modell trainieren, das Wölfe von Huskys unterscheiden kann. Dazu nutzen wir ganz viele Fotos von Wölfen und Huskys als Beispiele. KI-System 1 sollte nun aus den Fotos lernen, dass sich z. B. Fellfarbe, Ohrenform, Schwanz und Schnauze der Tiere unterscheiden. Wenn diese Muster erkannt werden, dann kann das KI-System später auch auf neuen Fotos Huskys und Wölfe gut unterscheiden. **Allerdings können wir dem System nicht einfach sagen, was die richtigen Muster sind.**

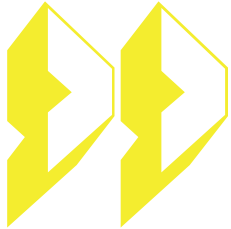
Sieh Dir die folgenden Bilder an und überlege, welches Muster ein KI-System hier stattdessen auch erkennen könnte:



Richtig! Alle Huskys sind im Schnee zu sehen – der Hintergrund ist also immer weiß – und alle Wölfe im Wald – der Hintergrund ist braun und grün. Wenn nun im Modell die Hintergrundfarbe als wichtiges Unterscheidungsmerkmal von Huskys und Wölfen aufgenommen wird, dann werden **alle neuen Bilder, die Huskys nicht im Schnee zeigen, als Wolf erkannt**. Das wäre natürlich nicht gut!



Das Modell gibt aus:
Wolf!



Wie lässt sich herausfinden, welche Merkmale ein KI-System in einem Bild benutzt, um zwei Gruppen zu unterscheiden?

Hier kommt **KI-System 2** ins Spiel: Es setzt eine **Methode der erklärbaren Künstlichen Intelligenz [eXplainable AI, kurz XAI]** um. Diese ermöglicht nachzuvollziehen, warum KI-System 1 ein bestimmtes Ergebnis liefert.

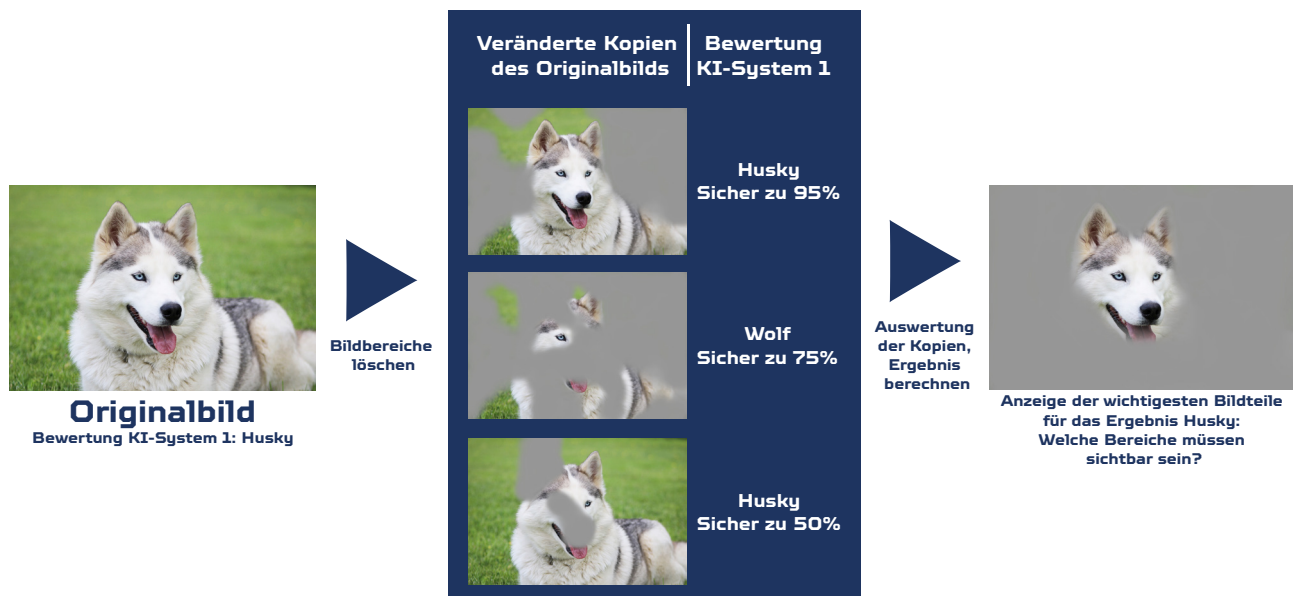
Dazu kann man zeigen oder beschreiben, welche Informationen wichtig für das Ergebnis waren. Es geht nicht darum, eine Erklärung zu geben, warum ein Bild **tatsächlich** zu einer bestimmten Gruppe gehört, also ein Husky oder Wolf ist, sondern es soll erklärt werden, warum das Modell diese Gruppe **errechnet** hat. **Dabei hilft das XAI-System, indem es beispielsweise bestimmt, welche Teile eines Bildes zu einem bestimmten Ergebnis beigetragen haben.**



Wie macht man das? Zum Beispiel mit der **LIME-Methode**: Von einem Bild werden viele Kopien erzeugt, aus denen verschiedene Teile gelöscht und durch graue Pixel ersetzt werden. Diese Bilder werden dann dem gelernten Modell (KI-System 1) gezeigt. Anschließend wird beobachtet, bei welchen veränderten Bildern das KI-System noch auf dasselbe Ergebnis kommt wie beim ursprünglichen Bild und bei welchen nicht. **Aus diesen Beobachtungen kann man dann berechnen, welche Bildbereiche wichtig waren und diese im ursprünglichen Foto markieren.** Durch die Markierung können die Entwickler und auch Du in der Aufgabe dann sehen, ob das System auf sinnvolle Bildmerkmale geachtet hat.



So arbeitet das XAI-System mit der LIME-Methode:



Es gibt viele Bereiche, in denen die Nachvollziehbarkeit, also warum ein KI-System zu einem bestimmten Ergebnis gekommen ist, wichtig ist. So ist es beispielsweise in der **Medizin** wichtig für Ärzt:innen zu verstehen, warum ein KI-System eine bestimmte Krankheit diagnostiziert hat. Erklärungen können auch dabei helfen zu **erkennen, ob das KI-System mit unfairen Daten trainiert wurde**, in denen z. B. verschiedene Geschlechter und Hautfarben nicht gut repräsentiert sind. Auch das kann in der Medizin problematisch sein, weil es dadurch zu falschen Diagnosen kommen kann.

Diese Lernstation wurde in Kooperation mit Ute Schmid, Lehrstuhl für Kognitive Systeme der Otto-Friedrich-Universität Bamberg, entwickelt.
Vielen Dank für die Zusammenarbeit!





QUELLEN

Infotext basierend auf:

Schmid, U. (2024): „Kapitel 9: Erklärbarkeit.“ in *Künstliche Intelligenz für Lehrkräfte. Eine fachliche Einführung mit didaktischen Hinweisen*, U. Furbach, E. Kitzelmann, T. Michaeli & U. Schmid (Hrsg.). Springer, 2024.

Fotos Huskys

<https://pixabay.com/de/photos/husky-hund-siberian-husky-2443664/>, Bild von Sonja Lindberg
<https://pixabay.com/de/photos/siberian-husky-schnee-hund-heiser-291721/>, Bild von forthdown
<https://pixabay.com/de/photos/hund-husky-tier-natur-haustier-4811796/>, Bild von Detroitius
<https://pixabay.com/de/photos/husky-schlittenfahrt-winter-schnee-2733209/>, Bild von Andreas

Fotos Wölfe

<https://pixabay.com/de/photos/wolf-raubtier-j%C3%A4ger-canis-lupus-635063/>,
 Bild von Rain Carnation
<https://pixabay.com/de/photos/wolf-wildtier-tierwelt-raubtier-3151876/>, Bild von Alexa
<https://pixabay.com/de/photos/wolf-predator-raubtier-1583200/>, Bild von Marcel Langthim
<https://pixabay.com/de/photos/kanada-omega-park-fauna-wei%C3%9Fer-wolf-4131643/>,
 Bild von LuisValiente

Foto „Das Modell gibt aus: Wolf!“

<https://pixabay.com/de/photos/husky-siberian-hund-schlittenhund-7729047/>, Bild von Jeannette1980

Graphik LIME-Methode

erstellt von Annabel Lindner

Graphik KI-System/XAI-System

erstellt von Annabel Lindner & Michaela Müller-Unterweger
 Laptop: <https://pixabay.com/de/vectors/laptop-pc-computer-informatik-2243898/>,
 Bild von FiveFlowersForFamilyFirst
 Lupe: <https://pixabay.com/de/vectors/lesebrille-erweiterung-lupe-1141525/>, Bild von SJ
 Pfeile: <https://pixabay.com/de/vectors/synchronisieren-pfeile-kreislauf-150123/>,
 Bild von OpenClipart-Vectors





BILDQUELLEN KLAPPKARTEN

Fotos Katzen

<https://pixabay.com/photos/animal-nature-wildlife-cat-mammals-3351691/>,

Bild von jumyoung youn

<https://pixabay.com/photos/animal-stray-cat-mammal-kitten-7740010/>, Bild von alicepaipai

Katze schwarz: Willi, Foto von Annabel Lindner

<https://pixabay.com/photos/cat-nature-predator-wildcat-feline-3594271/>,

Bild von Dewald Van Rensburg

<https://pixabay.com/photos/cat-domestic-cat-domestic-animal-4609833/>,

Bild von Andreas Glöckner

Perserkatze weiß/grau: Muffin, Foto von Annabel Lindner

Perserkatze schwarz: Cookie, Foto von Annabel Lindner

Tabby braun getigert: Miss Sophie, Foto von Annabel Lindner

<https://pixabay.com/photos/cat-sleeping-cat-feline-pet-animal-2605502/>,

Bild von Екатерина Гусева

<https://pixabay.com/de/photos/tier-katze-katzen-s%C3%A4ugetier-7105049/>,

Bild von mkluthke

<https://pixabay.com/de/photos/katze-tigerkatze-hauskatze-sprung-1817797/>, Bild von rihajj

<https://pixabay.com/de/photos/bengalkatze-katze-baum-haustier-6307542/>,

Bild von Jeannette1980

<https://pixabay.com/photos/scottish-wildcat-wildcat-feline-8232790/>, Bild von Angela

<https://pixabay.com/photos/lion-lioness-few-wildlife-big-cat-2845896/>, Bild von Jana V. M.

<https://pixabay.com/photos/wildlife-lynx-hunter-predator-4397295/>, Bild von G.C.

<https://pixabay.com/photos/sand-cat-wildcat-zoo-berlin-1278537/>, Bild von Hans Benn

<https://pixabay.com/photos/wildcats-young-animals-zoo-870643/>, Bild von Marcel Langthim

<https://pixabay.com/photos/tiger-grass-mammal-animal-cat-2463148/>, Bild von Stan Petersen

<https://pixabay.com/photos/lion-big-cat-predator-animal-mane-4357408/>,

Bild von Herbert Aust

<https://pixabay.com/photos/cheetah-africa-safari-cat-wildlife-7277665/>, Bild von Gareth Webb

<https://pixabay.com/photos/wildcat-africa-safari-wild-animal-1225323/>,

Bild von Poinger_Herzschlog

<https://pixabay.com/de/photos/schneeleopard-raubtier-raubkatze-4019923/>,

Bild von Dieter Staab

<https://pixabay.com/de/photos/l%C3%B6we-br%C3%BClle-afrika-tier-wildkatze-3012515/>,

Bild von Sarah Richter

<https://pixabay.com/photos/real-estate-tiger-fantasy-7132405/>, Bild von Anna Lisa



BILDQUELLEN KLAPPKARTEN

Fotos (nicht) traurige Menschen

<https://pixabay.com/de/photos/frau-weinen-tränen-traurig-1867127/>, Bild von Pexels
<https://pixabay.com/de/photos/mädchen-traurig-gesicht-modell-7735588/>,
 Bild von u_fqyx9glm41
<https://pixabay.com/de/photos/kind-mädchen-traurig-gesicht-6077904/>,
 Bild von briandastray03
<https://pixabay.com/de/photos/mädchen-traurig-gesicht-schöne-6553281/>, Bild von Nguyen Hai
<https://pixabay.com/photos/portrait-girl-body-retro-face-8441126/>, Bild von Alexander Krivitskiy
<https://pixabay.com/de/photos/frau-gesicht-sommersprossen-lächeln-1867431/>, Bild von Pexels
<https://pixabay.com/de/photos/mann-porträt-schwarz-weiss-gesicht-2940314/>, Bild von Peter H
<https://pixabay.com/de/photos/mann-verwirrt-jung-männlich-brille-5914348/>,
 Bild von Usman Yousaf
<https://pixabay.com/de/photos/mann-portrt-gesicht-vietnam-8946888/>,
 Bild von Mehrshad Rezaei
<https://pixabay.com/de/photos/kind-mädchen-blond-gesicht-blick-660695/>, Bild von Petra
<https://pixabay.com/de/photos/traurig-mädchen-gesicht-porträt-4687405/>,
 Bild von Chu Viet Đôn
<https://pixabay.com/de/photos/mann-traurig-einfarbig-gesicht-844212/>, Bild von Leroy Skalstad
<https://pixabay.com/de/photos/mann-traurig-dunkel-gesicht-915230/>, Bild von Suvajit Roy
<https://pixabay.com/de/photos/kind-traurig-trauriges-kind-1000610/>, Bild von Marko Lovric
<https://pixabay.com/de/photos/mädchen-besorgt-porträt-gesicht-738306/>,
 Bild von bess.hamiti@gmail.com
<https://pixabay.com/de/photos/junge-kind-zaun-traurig-529067/>, Bild von lisa runnels
<https://pixabay.com/de/photos/frau-modell-porträt-pose-regen-4644496/>, Bild von Nika Akin
<https://pixabay.com/de/photos/schönheit-schöne-frau-hotel-3625320/>,
 Bild von Rassameejantra
<https://pixabay.com/de/photos/junge-bad-nass-gesicht-lächeln-67852/>, Bild von lisa runnels
<https://pixabay.com/de/photos/emotionen-mann-glücklich-traurig-371238/>,
 Bild von Ryan McGuire, zugeschnitten
<https://pixabay.com/de/photos/kleines-mädchen-strand-traurig-7339106/>,
 Bild von Boris Gonzalez
<https://pixabay.com/de/photos/blau-mädchen-traurig-gesicht-kopf-1382940/>,
 Bild von Khusen Rustamov, Farbmodifikation
<https://pixabay.com/de/photos/gesicht-träne-weinen-traurig-5144358/>, Bild von inna mykytas
<https://www.pexels.com/de-de/foto/frauen-beziehung-jung-sweatshirt-6383157/>,
 Bild von Liza Summer