TRUST, BUT VERIFY!

AI systems that use machine learning are often „black boxes". As they have been trained on large amounts of data, it is often impossible to say exactly how such an AI system arrived at a result. This is a problem, because how can the results of AI systems then be verified? With other AI systems, of course! These use methods known as eXplainable Artificial Intelligence (or XAI for short).

The Problem:

With normal algorithms, we can follow all the instructions and therefore know exactly how programs calculate something. With AI algorithms such as neural networks (learn more about them in Station 03), this is no longer so simple: the network performs a complex calculation with a large number of numbers that represent all the information entered into the network. The result is an output to display.

However, it is no longer clear to humans which number represents which piece of information and how important it is. For this reason, even the developers of the AI system are not able to check whether an AI system has found the right patterns and rules in a data set, or whether it has learned something wrong. It can happen that the AI system works well on the training data, i.e. the examples it has used to learn, but no longer works on new inputs. This is known as overfitting.

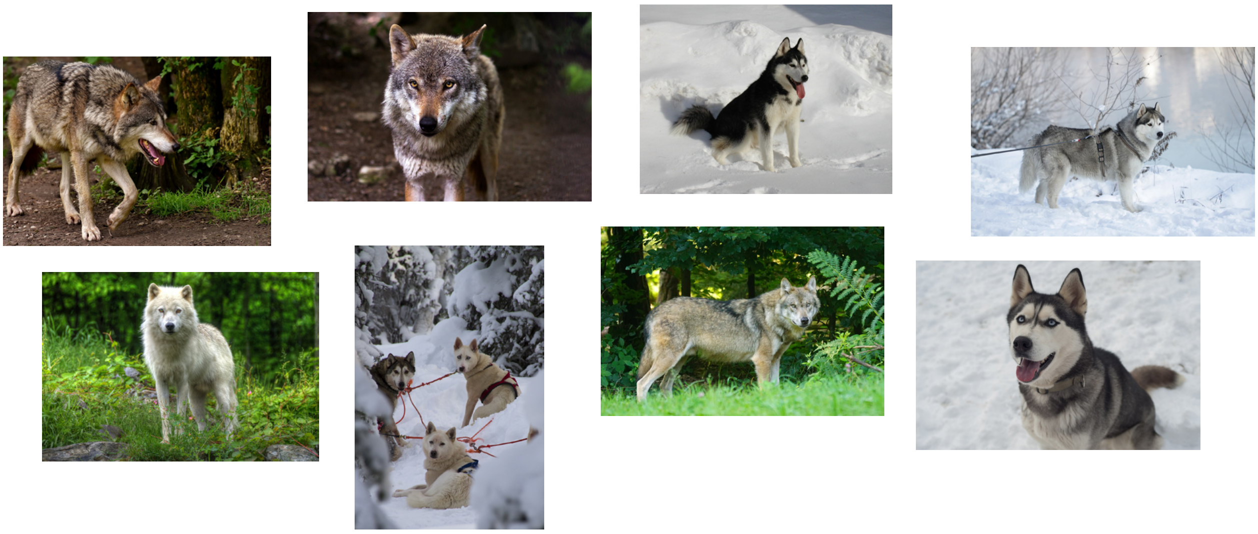
Another problem is that the training data is not balanced, but only very specific data is available for certain groups of people. For example, a data set of images of teenagers will often show light-skinned teenagers playing sports, while dark-skinned teenagers will often be shown getting arrested. This problem is called sampling bias (bias in the selection of training data) or unfair bias.

What does this mean?

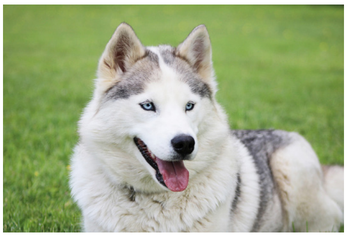
We work with two AI systems. **AI system 1** uses machine learning to build a model from data. This model is a black box, so it is not possible to determine exactly how it arrives at its result

We want to train a model that can distinguish wolves from huskies. To do this, we use many photos of wolves and huskies as examples. AI System 1 should learn from the photos that the animals have different fur colors, ear shapes, tails, and muzzles. By recognizing these patterns, the AI system will later be able to distinguish between huskies and wolves in new photos. However, we cannot simply tell the system what the correct patterns are.

Take a look at the following photos and think about what pattern an AI system might recognize instead:



That's right! Every husky is in the snow, so the background is always white, and every wolf is in the woods, so the background is brown and green. If the background color is now included in the model as an important distinguishing feature between huskies and wolves, then all new images that do not show huskies in the snow will be recognized as wolves. Obviously that would not be good!

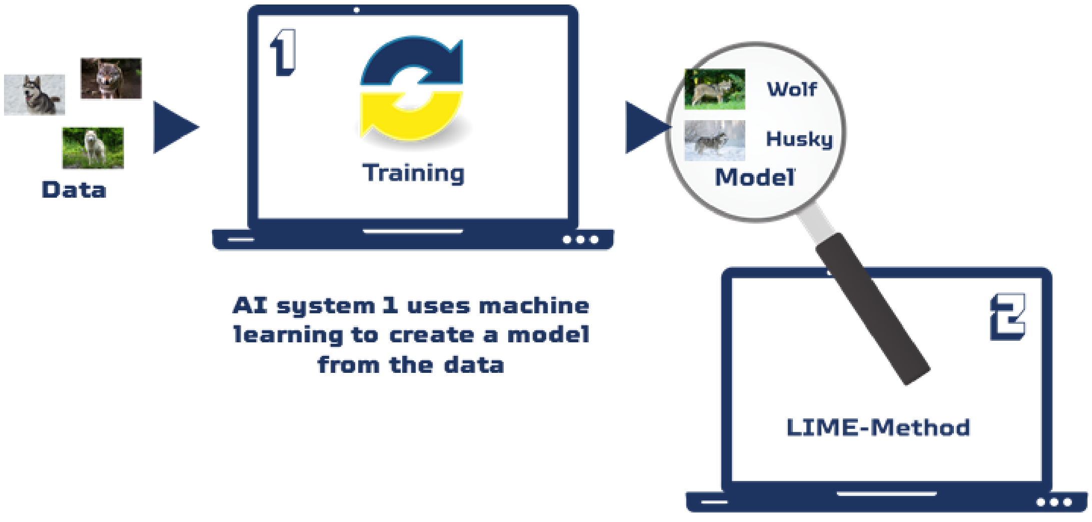


The output of the model: Wolf!

How can we find out what features an AI system uses in an image to distinguish between two groups?

This is where **AI System 2** comes in: it implements a method of eXplainable AI (or XAI for short). This makes it possible to understand why AI System 1 delivers a certain result.

You can show or describe which information was important for the result. The goal is not to explain why an image **actually** belongs to a certain group, i.e. why it is a husky or a wolf, but to explain why the model **calculated** this group. The XAI system helps with this by, for example, determining which parts of an image contributed to a particular result.

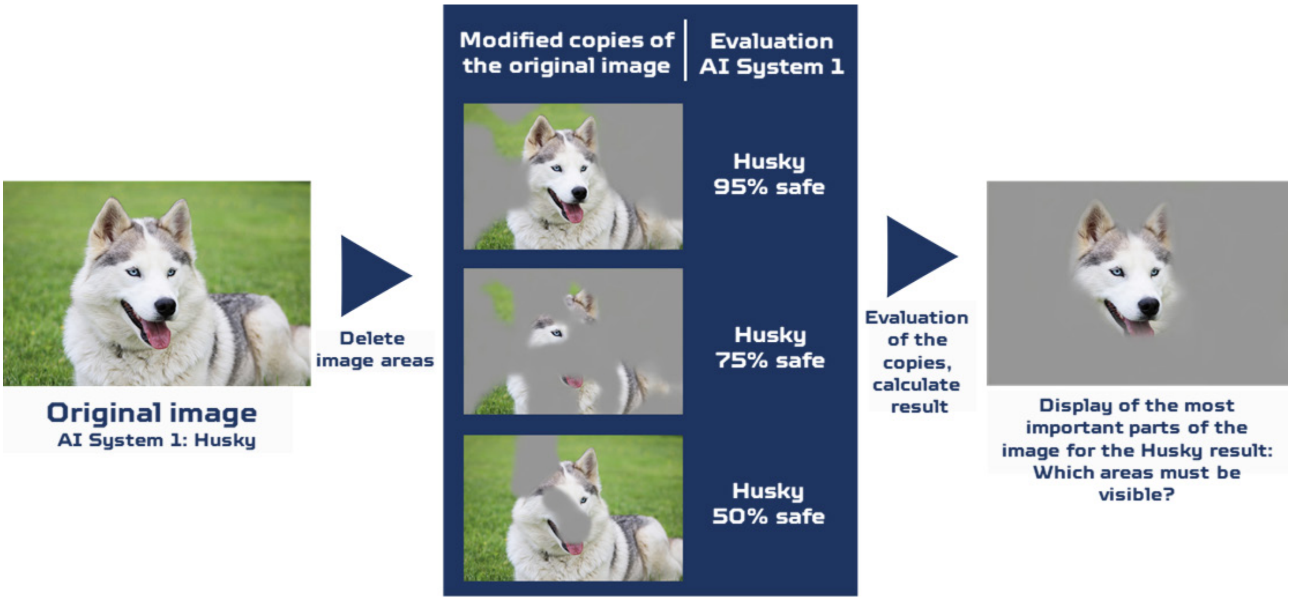


AI system 1 uses machine learning to create a model from the data

AI System 2 uses an XAI method to find the information that led to the result of AI System 1.

How do you do that? For example, with the LIME method: many copies of an image are made, from which various parts are deleted and replaced by gray pixels. These images are then shown to the trained model (AI system 1). It is then observed for which modified images the AI system still reaches the same result as the original image and for which it does not. These observations can then be used to calculate which areas of the image were important and mark them in the original photo. These markers allow you and the developers to see in the task whether the system paid attention to meaningful image features.

This is how the XAI system works with the LIME method:


Original image
AI System 1: Husky
Delete image areas
Modified copies of the original image
Evaluation AI System 1
Husky 95% safe
Husky 75% safe
Husky 50% safe
Evaluation of the copies, calculate result
Display of the most important parts of the image for the Husky result: Which areas must be visible?

There are many areas where traceability, or knowing why an AI system came to a certain conclusion, is important. In medicine, for example, it is important for doctors to understand why an AI system has diagnosed a particular disease. Explanations can also be an indicator of whether the AI system has been trained on inequitable data, where, for example, different genders and skin colors are not well represented. This can also be problematic in medicine, as it can lead to incorrect diagnoses.

This learning station was developed in cooperation with Ute Schmid, Chair for Cognitive Systems at the University of Bamberg. Many thanks for the collaboration!

SOURCES

Information text based on:

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.

Photos Huskies

https://pixabay.com/de/photos/husky-hund-siberian-husky-2443664/, Photo by Sonja Lindberg

https://pixabay.com/de/photos/siberian-husky-schnee-hund-heiser-291721/, Photo by forthdown

https://pixabay.com/de/photos/hund-husky-tier-natur-haustier-4811796/, Photo by Detroitius

https://pixabay.com/de/photos/husky-schlittenfahrt-winter-schnee-2733209/, Photo by Andreas

Photo Wolves

https://pixabay.com/de/photos/wolf-raubtier-j%C3%A4ger-canis-lupus-635063/, Photo by Rain Carnation

https://pixabay.com/de/photos/wolf-wildtier-tierwelt-raubtier-3151876/, Photo by Alexa

https://pixabay.com/de/photos/wolf-predator-raubtier-1583200/, Photo by Marcel Langthim

https://pixabay.com/de/photos/kanada-omega-park-fauna-wei%C3%9Fer-wolf-4131643/, Photo by LuisValiente

Photo "The output of the model: Wolf!"

https://pixabay.com/de/photos/husky-siberian-hund-schlittenhund-7729047/, Photo by Jeannette1980

Graphik LIME-Method

created by Annabel Lindner

Graphik KI-System/XAI-System

created by Annabel Lindner & Michaela Müller-Unterweger

Laptop: https://pixabay.com/de/vectors/laptop-pc-computer-informatik-2243898/, Photo by FiveFlowersForFamilyFirst

Magnifier: https://pixabay.com/de/vectors/lesebrille-erweiterung-lupe-1141525/, Photo by SJ

Arrow: https://pixabay.com/de/vectors/synchronisieren-pfeile-kreislauf-150123/, Photo by OpenClipart-Vectors

PICTURESOURCES FOLDING CARDS

Photos Cats

https://pixabay.com/photos/animal-nature-wildlife-cat-mammals-3351691/,
Photo by jumyoung youn
https://pixabay.com/photos/animal-stray-cat-mammal-kitten-7740010/, Photo by alicepaipai
black cat: Willi, Photo by Annabel Lindner
https://pixabay.com/photos/cat-nature-predator-wildcat-feline-3594271/,
Photo by Dewald Van Rensburg
https://pixabay.com/photos/cat-domestic-cat-domestic-animal-4609833/,
Photo by Andreas Glöckner
Persian cat white/grey: Muffin, Photo by Annabel Lindner
Persian cat black: Cookie, Photo by Annabel Lindner
TTabby brown mackerel: Miss Sophie, Photo by Annabel Lindner
https://pixabay.com/photos/cat-sleeping-cat-feline-pet-animal-2605502/,
Photo by Екатерина Гусева
https://pixabay.com/de/photos/tier-katze-katzen-s%C3%A4ugetier-7105049/,
Photo by mkluthke
https://pixabay.com/de/photos/katze-tigerkatze-hauskatze-sprung-1817797/, Photo by rihaij
https://pixabay.com/de/photos/bengalkatze-katze-baum-haustier-6307542/,
Photo by Jeannette1980
https://pixabay.com/photos/scottish-wildcat-wildcat-feline-8232790/, Photo by Angela
https://pixabay.com/photos/lion-lioness-few-wildlife-big-cat-2845896/ , Photo by Jana V. M.
https://pixabay.com/photos/wildlife-lynx-hunter-predator-4397295/, Photo by G.C.
https://pixabay.com/photos/sand-cat-wildcat-zoo-berlin-1278537/, Photo by Hans Benn
https://pixabay.com/photos/wildcats-young-animals-zoo-870643/, Photo by Marcel Langthim
https://pixabay.com/photos/tiger-grass-mammal-animal-cat-2463148/, Photo by Stan Petersen
https://pixabay.com/photos/lion-big-cat-predator-animal-mane-4357408/,
Photo by Herbert Aust
https://pixabay.com/photos/cheetah-africa-safari-cat-wildlife-7277665/, Photo by Gareth Webb
https://pixabay.com/photos/wildcat-africa-safari-wild-animal-1225323/,
Photo by Poinger_Herzschlog
https://pixabay.com/de/photos/schneeleopard-raubtier-raubkatze-4019923/,
Photo by Dieter Staab
https://pixabay.com/de/photos/l%C3%B6we-br%C3%BCllen-afrika-tier-wildkatze-3012515/,
Photo by Sarah Richter
https://pixabay.com/photos/real-estate-tiger-fantasy-7132405/, Photo by Anna Lisa

PICTURESOURCES FOLDING CARDS

Photos XY

https://pixabay.com/photos/animal-nature-wildlife-cat-mammals-3351691/,
Photo by jumyoung youn