

Was kann Künstliche Intelligenz leisten?

CAS Digitale Technologien und Innovation, 13. September 2019



Thilo Stadelmann

A photograph of a forest with tall, thin trees. Sunlight filters through the canopy in bright, golden rays, creating a dramatic play of light and shadow on the forest floor. The overall atmosphere is misty and serene.

Was ist KI?
Warum ist das jetzt aktuell?
Wie funktioniert das?



datalab
www.zhaw.ch/datalab

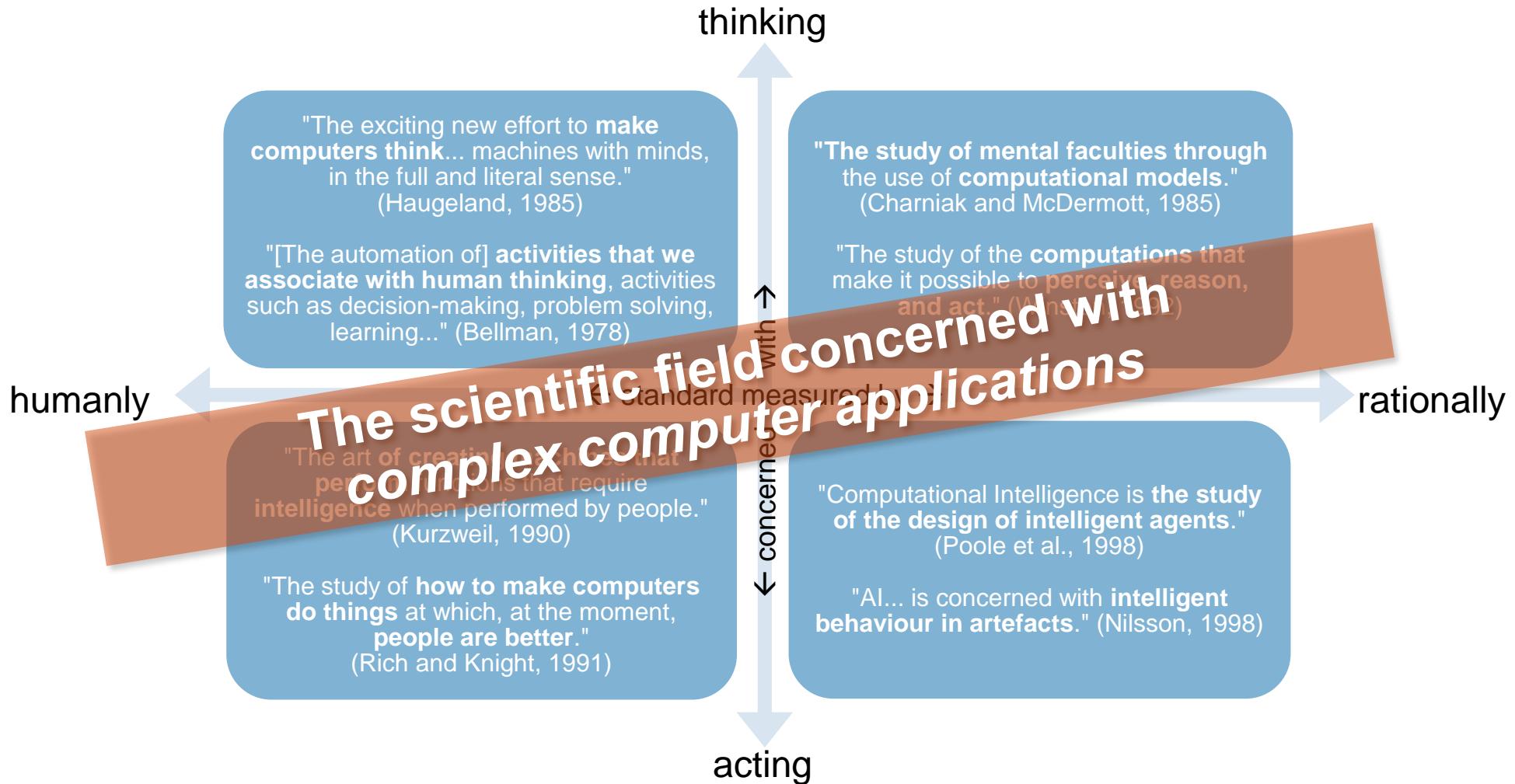
Was → Warum? → Wie?



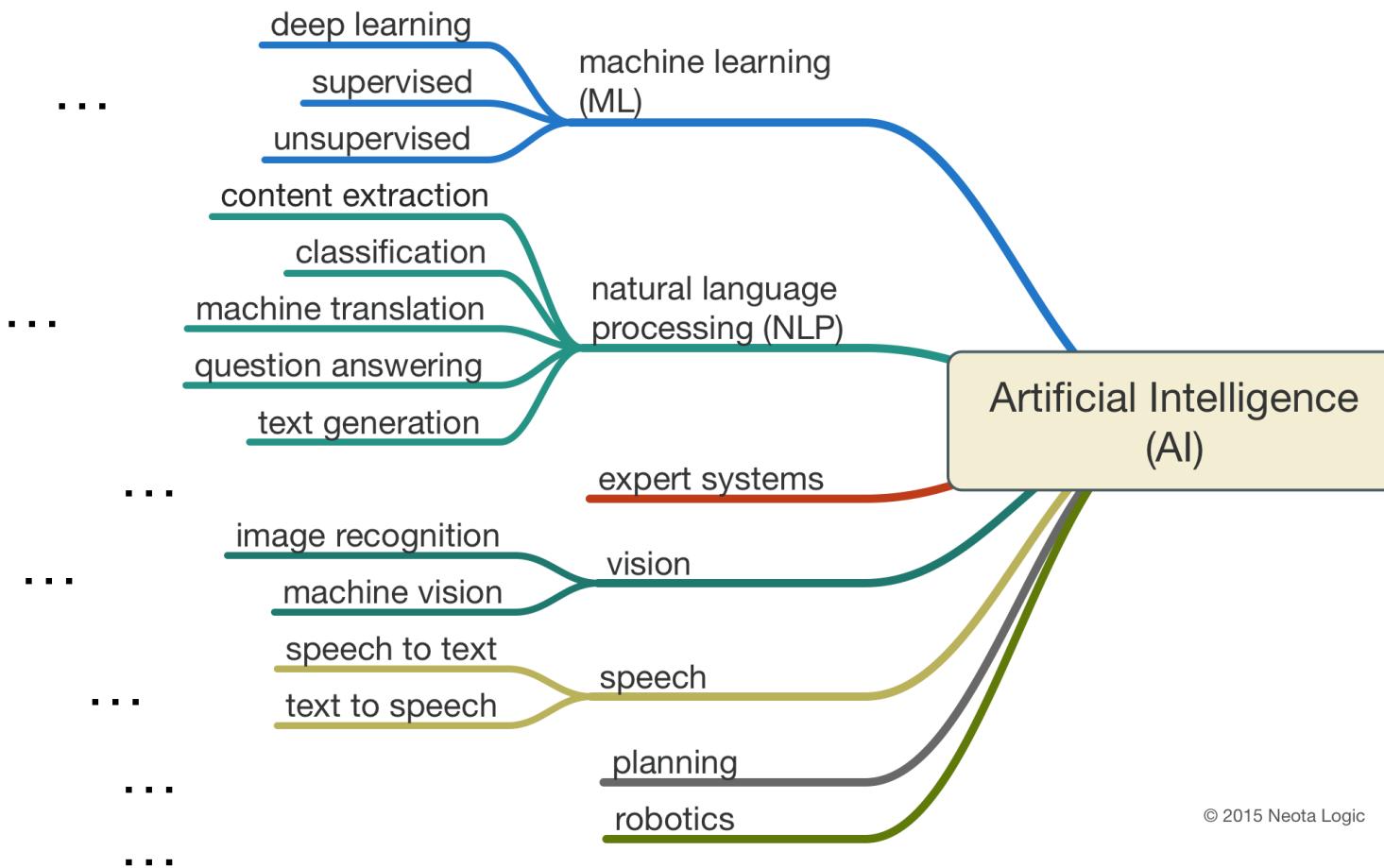
1

Was ist Künstliche Intelligenz?

Was ist künstliche Intelligenz?

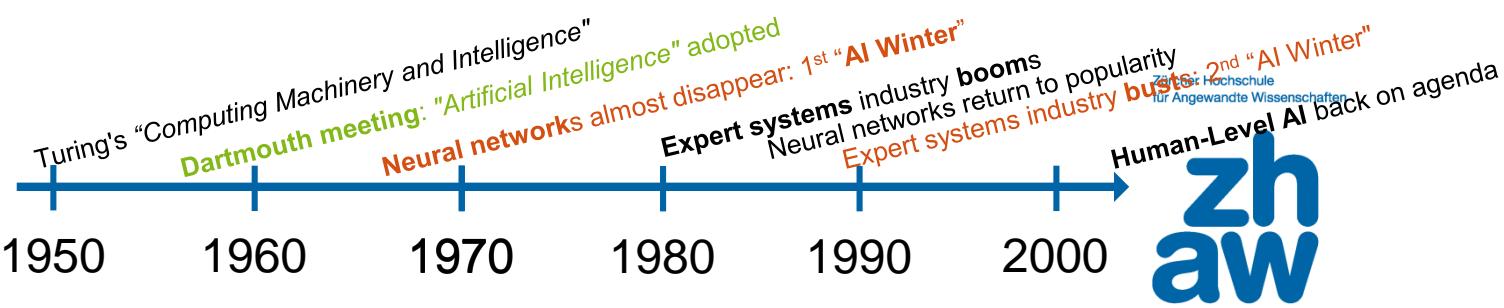


Was gehört zu künstlicher Intelligenz?



© 2015 Neota Logic

KI im Kontext



zhaw
Zürcher Hochschule
für Angewandte Wissenschaften



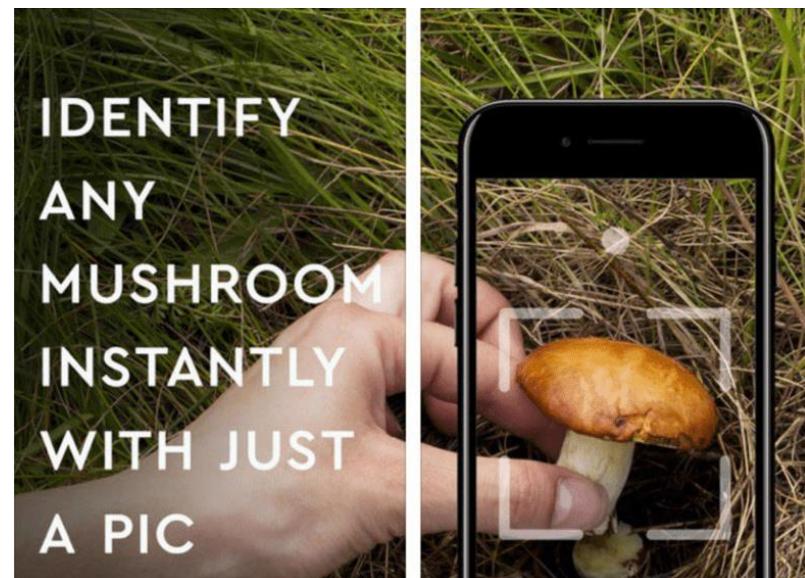
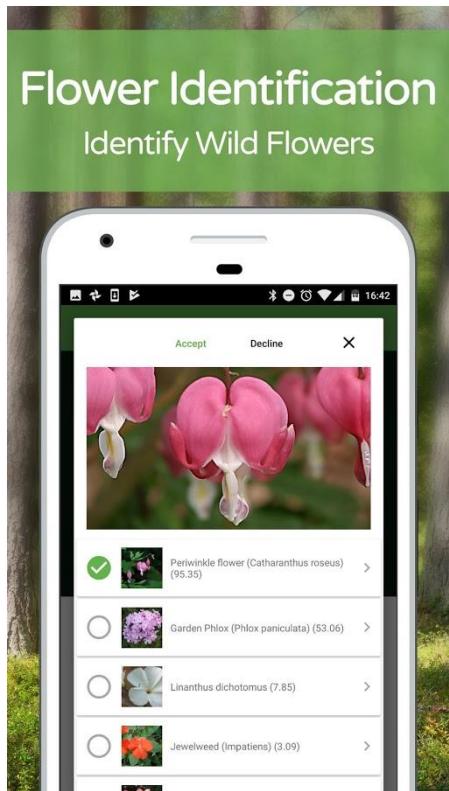
Was kann KI bereits heute?

- | | |
|---|--------------------------|
| 1. Play a decent game of table tennis | ok |
| 2. Drive safely along a curving mountain road | ok |
| 3. Drive safely along Technikumstrasse Winterthur | ok (only since recently) |
| 4. Buy a week's worth of groceries on the web | ok |
| 5. Buy a week's worth of groceries at Migros | no |
| 6. Play a decent game of bridge | ok |
| 7. Discover and prove a new mathematical theorem | not completed |
| 8. Design and execute a research program in molecular biology | not completed |
| 9. Write an intentionally funny story | no |
| 10. Give competent legal advice in a specialized area of law | ok |
| 11. Translate spoken English into spoken Swedish in real time | ok |
| 12. Converse successfully with another person for an hour | no |
| 13. Perform a complex surgical operation | not completed |
| 14. Unload any dishwasher and put everything away | no |
| 15. Compete in the game show Jeopardy! | ok |
| 16. Write clickbait articles fully automatized | ok |



Beispiel: Machbar vs. gefährlich

Technologie: Computer Vision mit Deep Learning



<https://www.cultofmac.com/495088/avoid-potentially-deadly-ai-app/>

Beispiel: Markterfolg vs. regulatorische Hürden

Technologie: Recommender Systems

Customers Who Bought This Item Also Bought



**Reckoning with Risk:
Learning to Live with Uncertainty**
by Gerd Gigerenzer
★★★★★ (8) £6.49



**Gut Feelings: The
Intelligence of the
Unconscious** by Gerd
Gigerenzer
£10.27



**Bounded Rationality: The
Adaptive Toolbox** by
Gerd Gigerenzer
£20.95

What Do Customers Ultimately Buy After Viewing This Item?



**68% buy
Simple Heuristics That Make Us Smart (Evolution & Cognition)**
£18.99



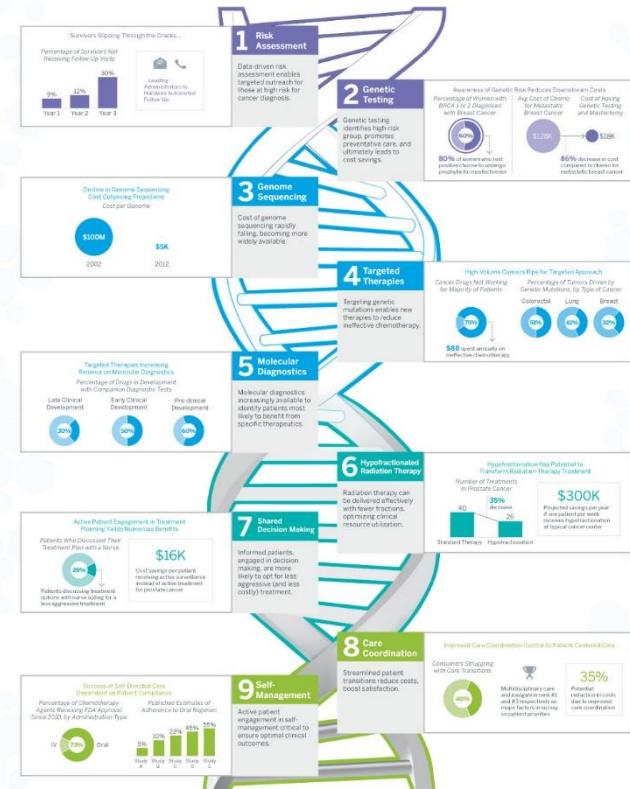
**17% buy
Gut Feelings: Short Cuts to Better Decision Making**
£6.74



**9% buy
Influence: The Psychology of Persuasion** ★★★★☆ (12)
£7.09

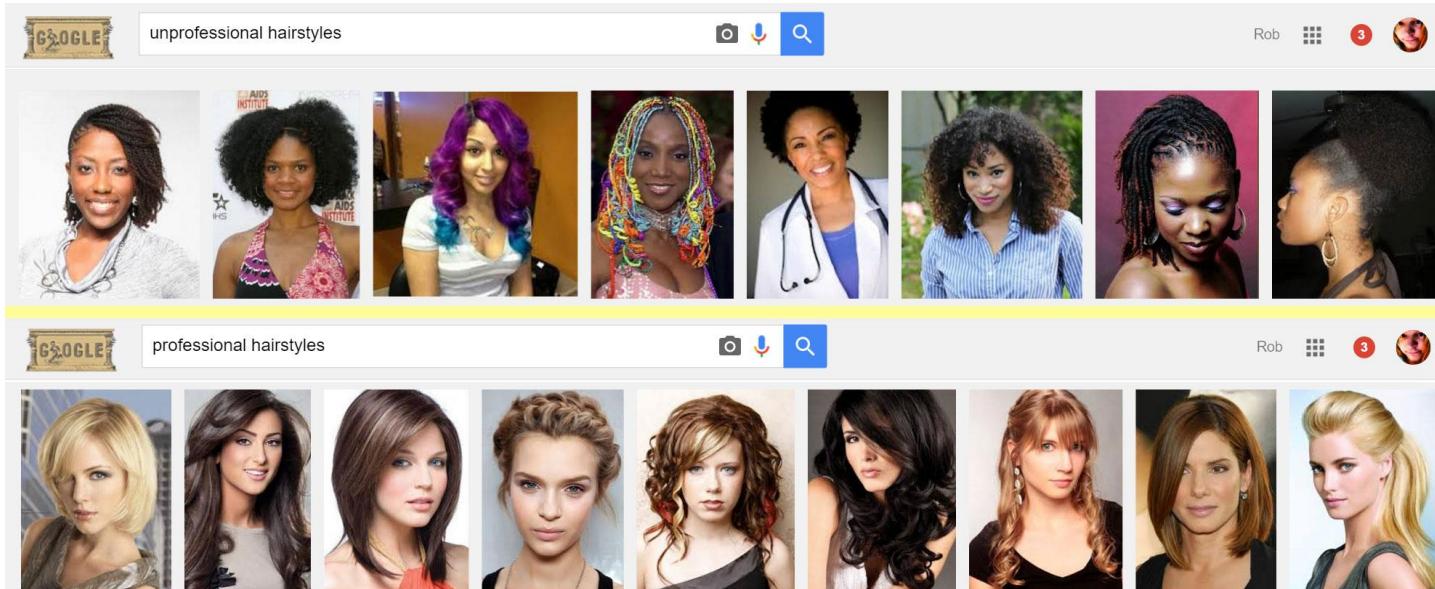
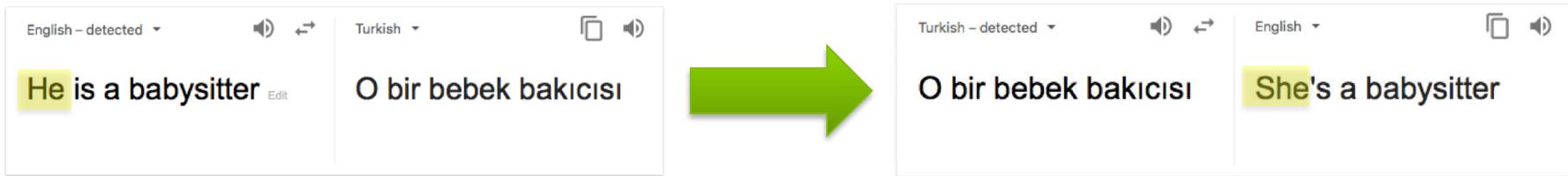
The Journey to Personalized Medicine

After years of anticipation, clinical innovations will soon make personalized medicine widely available. However, to realize its promise, providers will need to integrate clinical innovations with care delivery redesign.



Beispiel: Statistik vs. Bias

Technologie: Machine Learning



See also: Nassim Nicholas Taleb, «*The Black Swan: The Impact of the Highly Improbable*», 2007

Beispiel: künstl. Intelligenz vs. natürl. Dummheit

Technologie: Machine Learning mit nachgelagerten Regeln

The screenshot shows a blog post from the website 'SKYLIGHT'. The header features a dark bar with the 'SKYLIGHT' logo and navigation links for 'ABOUT US', 'SERVICES', and 'BLOG'. Below the header is a large, artistic photograph of a human brain with orange and red particles or dust floating around it. The main title of the post is 'Cylance, I Kill You!', dated '18 July 2019'. Below the title are social media sharing icons for Twitter, LinkedIn, Facebook, and Reddit. A brief summary reads: 'Read about our Journey of dissecting the brain of a leading AI based Endpoint Protection Product, culminating in the creation of a universal bypass'. The 'TL;DR' section explains that AI applications in security are clear and potentially useful, but AI-based products offer a new attack surface due to their internal model and feature extraction process. It highlights a specific bias in Cylance's AI-based antivirus product towards a game, which was exploited to create a universal bypass. The post concludes by stating that this method was successful for 100% of top 10 malware in May 2019 and approximately 90% for a larger sample of 384 malware.

18 July 2019

Cylance, I Kill You!

Read about our Journey of dissecting the brain of a leading AI based Endpoint Protection Product, culminating in the creation of a universal bypass

TL;DR

AI applications in security are clear and potentially useful, however AI based products offer a new and unique attack surface. Namely, if you could truly understand how a certain model works, and the type of features it uses to reach a decision, you would have the potential to fool it consistently, creating a universal bypass.

By carefully analyzing the engine and model of Cylance's AI based antivirus product, we identify a peculiar bias towards a specific game. Combining an analysis of the feature extraction process, its heavy reliance on strings, and its strong bias for this specific game, we are capable of crafting a simple and rather amusing bypass. Namely, by appending a selected list of strings to a malicious file, we are capable of changing its score significantly, avoiding detection. This method proved successful for 100% of the top 10 Malware for May 2019, and close to 90% for a larger sample of 384 malware.

Gefahren durch KI?

- KI ist per Definition eine “**dual use Technology**”
→ siehe Report von Brundage et al., 2018
- Aber: “**natürliche Dummheit**” ist die grössere Bedrohung
- **Algorithmische Ethik** und **erklärbare KI** sind in den letzten Jahren zu einem top Forschungsfeld geworden – nicht wegen der unkalkulierbaren Risiken per se, sondern:



A dark grey rectangular document cover. At the top, it lists several organizations: Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation, and OpenAI. Below this, the title "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation" is centered in a large, white, sans-serif font. A small "February 2018" is at the bottom right. The bottom half of the cover features a grid pattern of white symbols: dashes, slashes, dots, and plus signs.

Was → Warum? → Wie?



2

Warum ist das jetzt aktuell?
(Eine kurze Geschichte der letzten Jahre)

Google Acquires Artificial Intelligence Startup DeepMind For More Than \$500M

Posted Jan 26, 2014 by Catherine Shu (@catherineshu)



The graph plots Elo Rating against time in days. The Y-axis ranges from -2000 to 5000. The X-axis ranges from 0 to 40. Three series are shown: AlphaGo Zero 40 blocks (blue line), AlphaGo Lee (green dots), and AlphaGo Master (blue dots). AlphaGo Zero 40 blocks starts at -2000 and rises sharply to about 4800 by day 10, then continues to rise more slowly to approximately 5100 by day 40. AlphaGo Lee and AlphaGo Master are positioned at higher Elo ratings, around 4500 and 4800 respectively, and remain relatively flat.

40 days

AlphaGo Zero surpasses all other versions of AlphaGo and, arguably, becomes the best Go player in the world. It does this entirely from self-play, with no human intervention and using no historical data.

— AlphaGo Zero 40 blocks •••• AlphaGo Lee •••• AlphaGo Master

Google will buy reports that they in talks to buy couldn't disclose deal terms.

The acquisition was originally confirmed by Google to Re/code.

At last – a computer program that can beat a champion Go player **PAGE 484**

ALL SYSTEMS GO

CONSERVATION
SONGBIRDS A LA CARTE
Illegal harvest of millions of Mediterranean birds
PAGE 452

RESEARCH ETHICS
SAFEGUARD TRANSPARENCY
Don't let openness backfire on individuals
PAGE 459

POPULAR SCIENCE
WHEN GENES GOT 'SELFISH'
Dawkins's calling card forty years on
PAGE 462

NATURE.COM/NATURE
28 January 2016 410
Vol. 529 No. 7587

047



Deep neural networks can now transfer the style of one photo onto another

And the results are impressive

by James Vincent | @jvincent | Mar 30, 2017, 1:53pm EDT

f SHARE

tweet TWEET

in LINKEDIN

Computing

Algorithm
Artistic
Other In

A deep neural n
other images.

by Emerging Tech

The nature of art
of Vincent Van Gogh
Edvard Munch's
humans recogni



Original photo

Reference photo

Result

You've probably heard of an AI technique known as "style transfer" — or, if you haven't heard of it, you've seen it. The process uses neural networks to apply the look and feel of one image to another, and appears in apps like [Prisma](#) and [Facebook](#). These style transfers, however, are stylistic, not photorealistic. They look good because they look like they've been painted. Now a group of researchers from Cornell University and Adobe have augmented

Ad closed by Google

Report this ad

AdChoices >

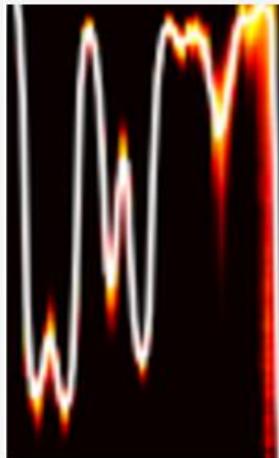


NOW TRENDING

WaveNet lässt Computergesproche natürlich klingen

von Henning Steier / 12.9.2017

Die Google-Tochter DeepMind macht auch Musik.



DeepMind lässt WaveNet Sprachsynthesen

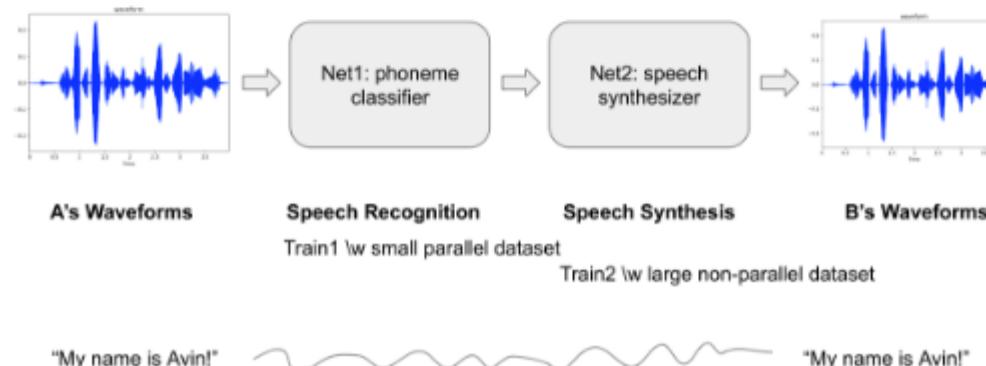
Die Google-Tochter DeepMind hat ein Spiel «Go» Schlagzeilen: einer der besten menschlichen Spieler schlägt einen der besten menschlichen Spieler. Londoner Unternehmen erzeugt Sprache, die sehr gut klingt. Sie erzeugt Sprache, die sehr gut klingt. im Blogeintrag des Unternehmens wird der Prozess dargestellt. Man kann den Massstab nimmt. Man hat

What if you could imitate a famous celebrity's voice or sing like a famous singer? This project started with a goal to convert someone's voice to a specific target voice. So called, it's voice style transfer. We worked on this project that aims to convert someone's voice to a famous English actress [Kate Winslet's voice](#). We implemented a deep neural networks to achieve that and more than 2 hours of audio book sentences read by Kate Winslet are used as a dataset.



Model Architecture

This is a many-to-one voice conversion system. The main significance of this work is that we could generate a target speaker's utterances without parallel data like <source's wav, target's wav>, <wav, text> or <wav, phone>, but only waveforms of the target speaker. (To make these parallel datasets needs a lot of effort.) All we need in this project is a number of waveforms of the target speaker's utterances and only a small set of <wav, phone> pairs from a number of anonymous speakers.



nerierte Sprache
is Texteingabe»

nerierte Musik
ine Inhaltsvorgabe»

...und die Liste liesse sich fortsetzen!

 Brandon Amos About Blog

Image Completion with Deep Learning in TensorFlow

August 9, 2016



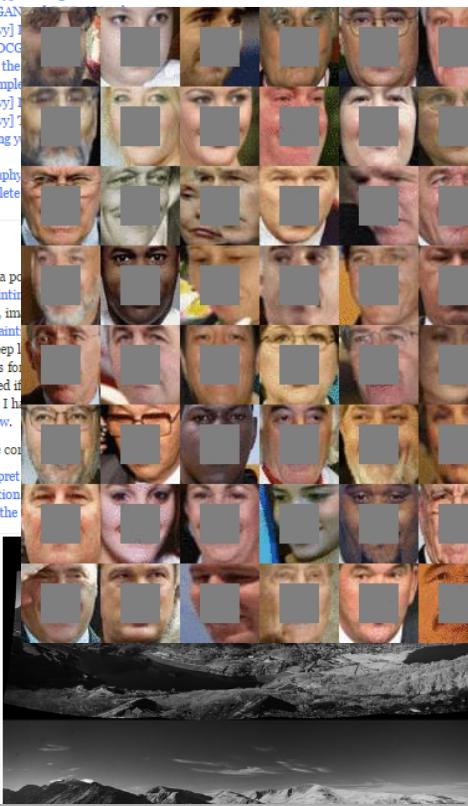
- Introduction
- Step 1: Interpreting images as samples from a probability distribution
 - How would you fill in the missing information?
 - But where does statistics fit in? These are images.
 - So how can we complete images?
- Step 2: Quickly generating fake images
 - Learning to generate new samples from an unknown probability distribution
 - [ML-Heavy] Generative Adversarial Net (GAN) building blocks
 - Using $G(z)$ to produce fake images
 - [ML-Heavy] Training DCGANs
 - Existing GANs
 - [ML-Heavy] DCGANs
 - Running DCGANs
- Step 3: Finding the right samples
 - Image completion
 - [ML-Heavy] Pseudo-GANs
 - [ML-Heavy] 1
 - [ML-Heavy] 2
 - Completing your own images
- Conclusion
- Partial bibliography
- Bonus: Incomplete

Introduction

Content-aware fill is a powerful technique for image completion and inpainting. It's great for filling in missing parts of images, but what if you want to do content-aware fill, image completion, and inpainting all at once? That's what "Semantic Image Inpainting" does. It shows how to use deep learning to fill in some deeper portions of images. This section can be skipped if you're not interested in learning about image completion or tensorflows.

We'll approach image completion by first

1. We'll first interpret the image
2. This interpretation will help us
3. Then we'll find the right samples





Nvidia AI Generates Fake Faces Based On Real Celebs

BY STEPHANIE MLOT 10.21.2017 :: 10:00AM EST

32 SHARES



I'm getting a distinctly mid-90s "The Rachel" vibe from the woman in the top left corner (via Nvidia)

STAY ON TARGET

AI Shelley Pens Truly Creepy Horror Stories—And You Can Help

Neural Network Serves Up Truly Frightening Halloween Costume Ideas

Celebrity scandals are about to get a lot more complicated.

Nvidia has developed a way of producing photo-quality, AI-generated human profiles—by using famous faces.

 Andrij Karpathy blog

About Hacker's guide to Neural Networks

The Unreasonable Effectiveness of Recurrent Neural Networks



the morning paper

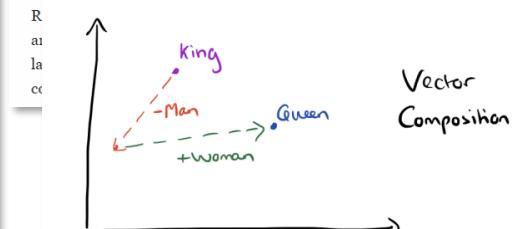
The amazing power of word vectors

APRIL 21, 2016

For today's post, I've drawn material not just from one paper, but from five! The subject matter is 'word2vec' – the work of Mikolov et al. at Google on efficient vector representations of words (and what you can do with them). The papers are:

- ★ Efficient Estimation of Word Representations in Vector Space – Mikolov et al. 2013
- ★ Distributed Representations of Words and Phrases and their Compositionality – Mikolov et al. 2013
- ★ Linguistic Regularities in Continuous Space Word Representations – Mikolov et al. 2013
- ★ word2vec Parameter Learning Explained – Rong 2014
- ★ word2vec Explained: Deriving Mikolov et al.'s Negative Sampling Word-Embedding Method – Goldberg and Levy 2014

From the first of these papers ('Efficient estimation...') we get a description of the *Continuous Bag-of-Words* and *Continuous Skip-gram* models for learning word vectors (we'll talk about what a word vector is in a moment...). From the second paper we get more illustrations of the power of word vectors, some additional information on optimisations for the skip-gram model (hierarchical softmax and negative sampling), and a discussion of *analogies* and *metaphors*.

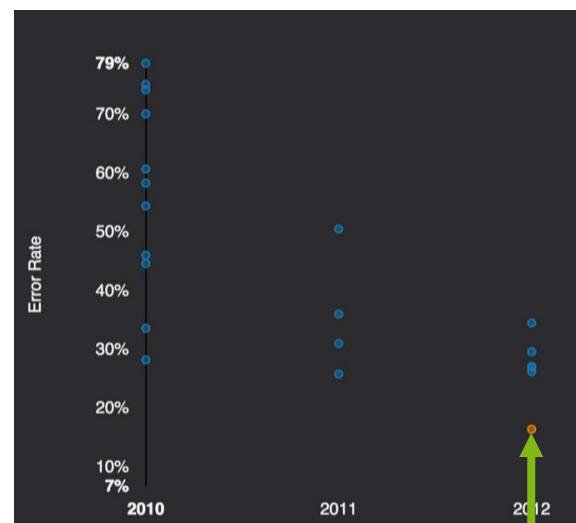


Was ist passiert?

Der ImageNet Wettbewerb



1000 Kategorien
1 Mio. Beispiele



A. Krizhevsky verwendet als erster ein
sog. «Deep Neural Network» (CNN)

2015: Computer haben "Sehen" gelernt

4.95% Microsoft (06. Februar)
→ Besser als Menschen (5.10%)

4.80% Google (11. Februar)

4.58% Baidu (11. Mai)

3.57% Microsoft (10. Dezember)

Was → Warum? → Wie?



3

Wie geht das?

Idee: Mehr «Tiefe» um Merkmale automatisch zu lernen

Classical image processing



Feature extraction
(SIFT, SURF, LBP, HOG, etc.)

(0.2, 0.4, ...)

Classification
(SVM, neural network, etc.)



Container ship

Automatisierung komplexer Prozesse basierend auf
(hoch-dimensionalem) Sensor-Input

Grundlage

Induktives überwachtes Lernen

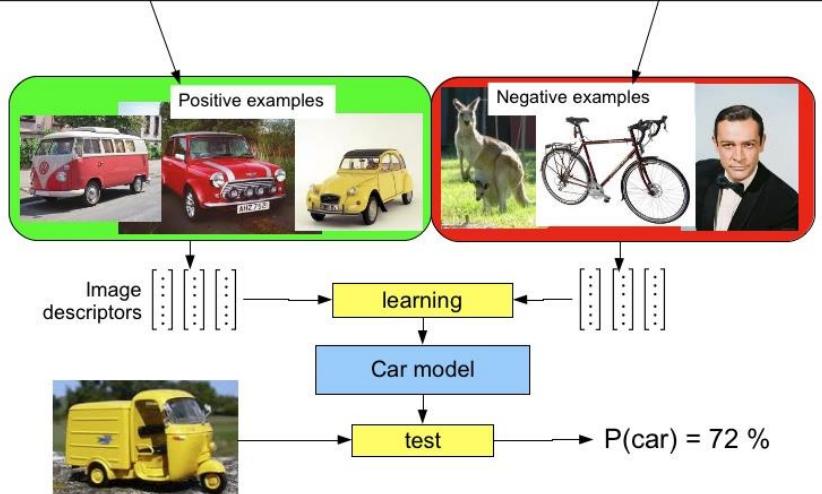
Annahme

- Ein an *genügend viele* Beispiele angepasstes Modell...
- ...wird auch auf unbekannte Daten **generalisieren**

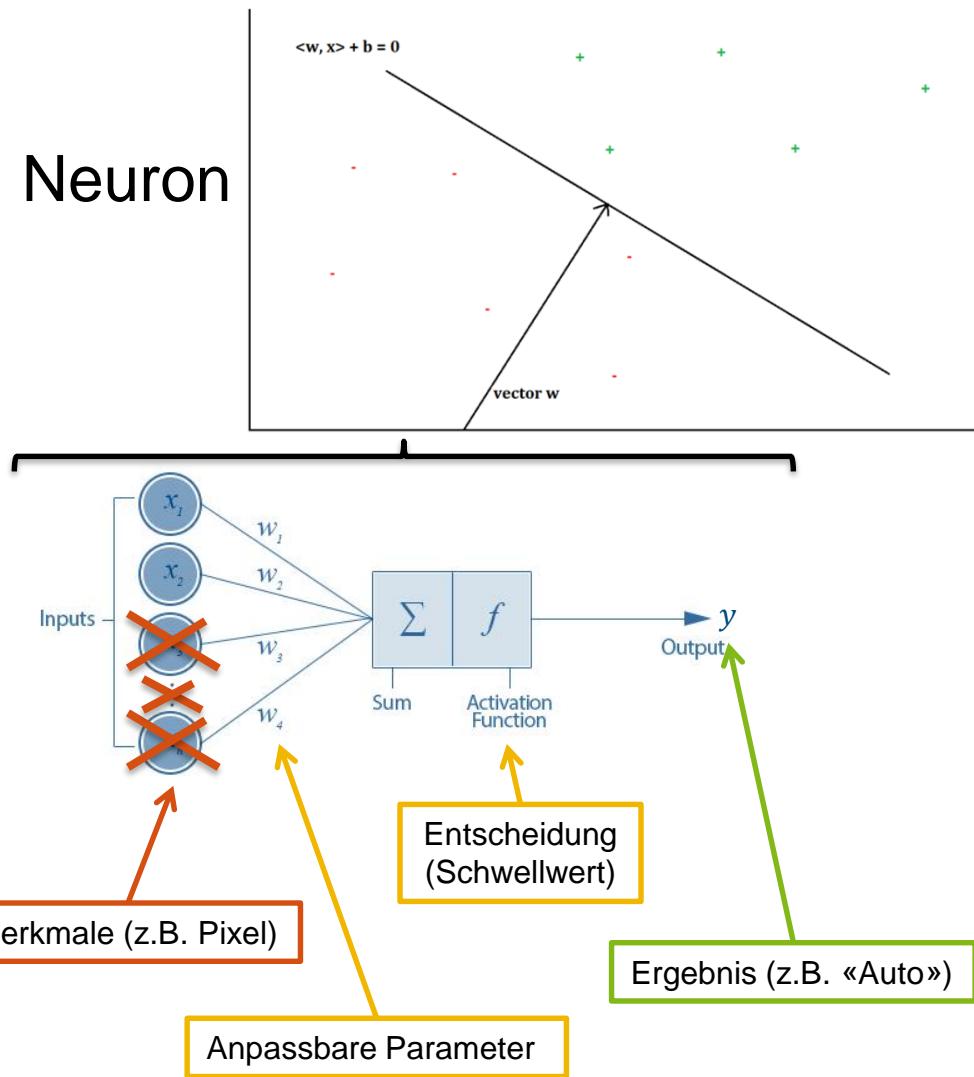
Methode

- **Suchen der Parameter einer gegebenen Funktion...**
- ...so dass für alle Beispiele Eingabe (Bild) auf Ausgabe («Auto») abgebildet wird

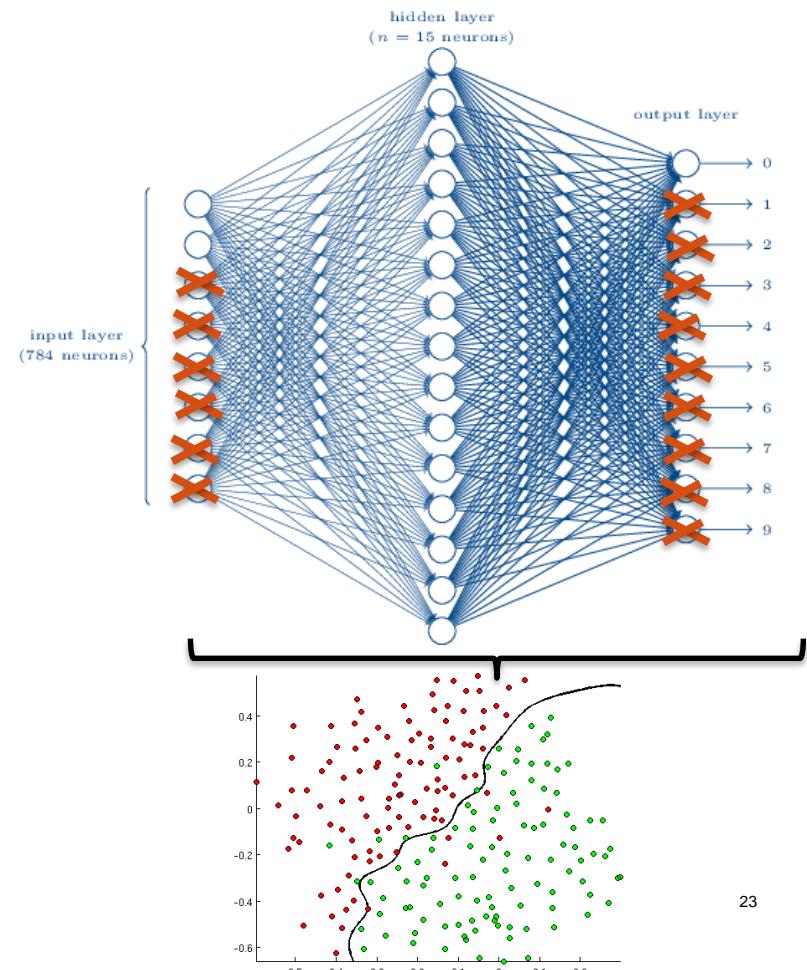
$$f(x) = y$$



Suche der Parameter einer Funktion?



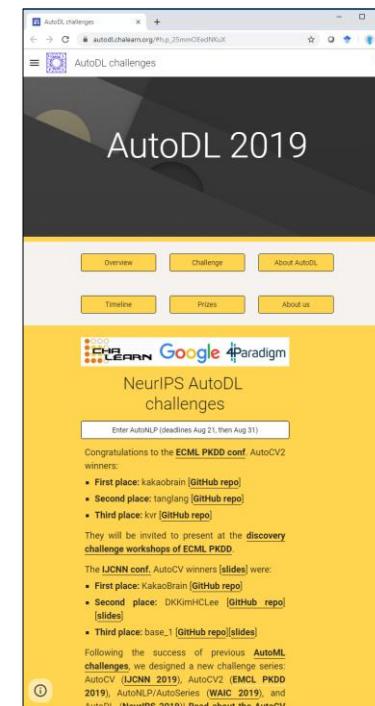
Neuronales Netz



Ausblick: Innovation aus neuronalen Netzen?



- *Deep Learning* ist der Kern vieler erstaunlicher Anwendungen im Bereich der Automatisierung von Wahrnehmung (Synthese & Analyse)
→ Code & Datensätze laden “**Maker**” zum Ausprobieren ein
- *Automated Deep Learning* macht **in Zukunft** vielleicht manche Anwendung einfacher
- Aktuell braucht es ein gutes Verständnis der Details der Methode, um kreativ neue Szenarien zu erdenken (Stichwort “**mTrainer**” anstatt Programmierer?)



Schlussfolgerungen



- Deep Learning hat zu Paradigmenwechsel in *Mustererkennungsaufgaben* geführt
- Die Zeit vom Grundlagenresultat zur praktischer Anwendung beträgt wenige Monate
- Es gibt Methoden zum Hineinschauen in neuronale Black Boxes (siehe Anhang)
- Spezifische Aufgaben lassen sich sehr gut automatisieren (z.B. Ähnlichkeitssuche)



Zu mir:

- Prof. KI/ML, Scientific Director ZHAW digital
- Email: stdm@zhaw.ch
- Telefon: 058 934 72 08
- Web: <https://stdm.github.io/>
- Twitter: @thilo_on_data
- LinkedIn: thilo-stadelmann



Mehr zum Thema:

- Data+Service Alliance: www.data-service-alliance.ch
- KI: <https://sgaico.swissinformatics.org/>
- Zusammenarbeit: datalab@zhaw.ch



ANHANG

Developing for algorithmic fairness

The FAT ML code of conduct

See <http://www.fatml.org/resources/principles-for-accountable-algorithms>

FAT / ML



Purpose

- Help developers to **build algorithmic systems in publicly accountable ways**
- Accountability: the **obligation to report, explain, or justify** algorithmic decision-making & **mitigate any negative social impacts** or potential harms

Premise

- *A human ultimately responsible for decisions made/informed by an algorithm*

Principles

- **Responsibility, Explainability, Accuracy, Auditability, Fairness**

Make available somebody who will take care of adverse individual / societal effects

Explain any **algorithmic decision** in non-technical terms to end users

Report all sources of uncertainty / error in algorithms & data

Enable 3rd parties to **probe & understand** system behavior

Ensure algorithmic **decisions are not discriminatory** w.r.t. to people groups

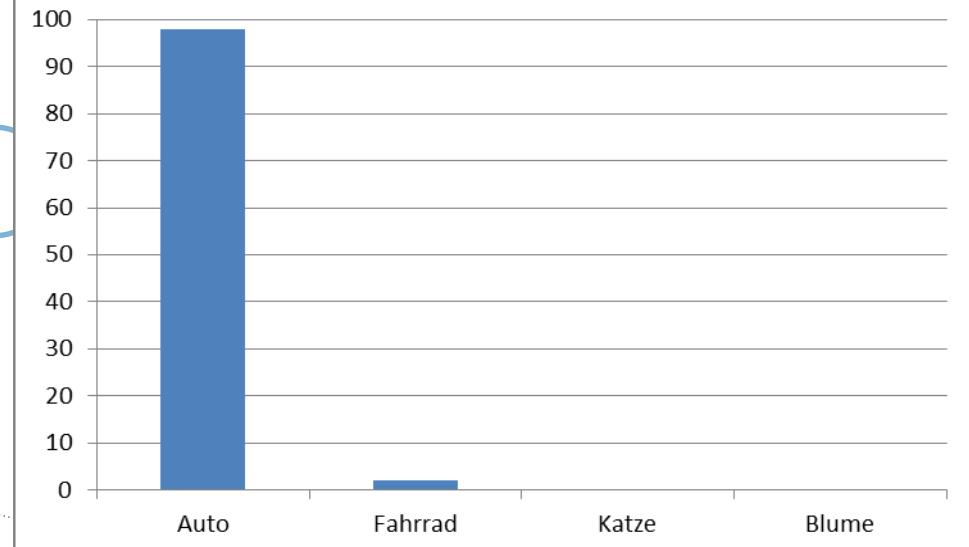
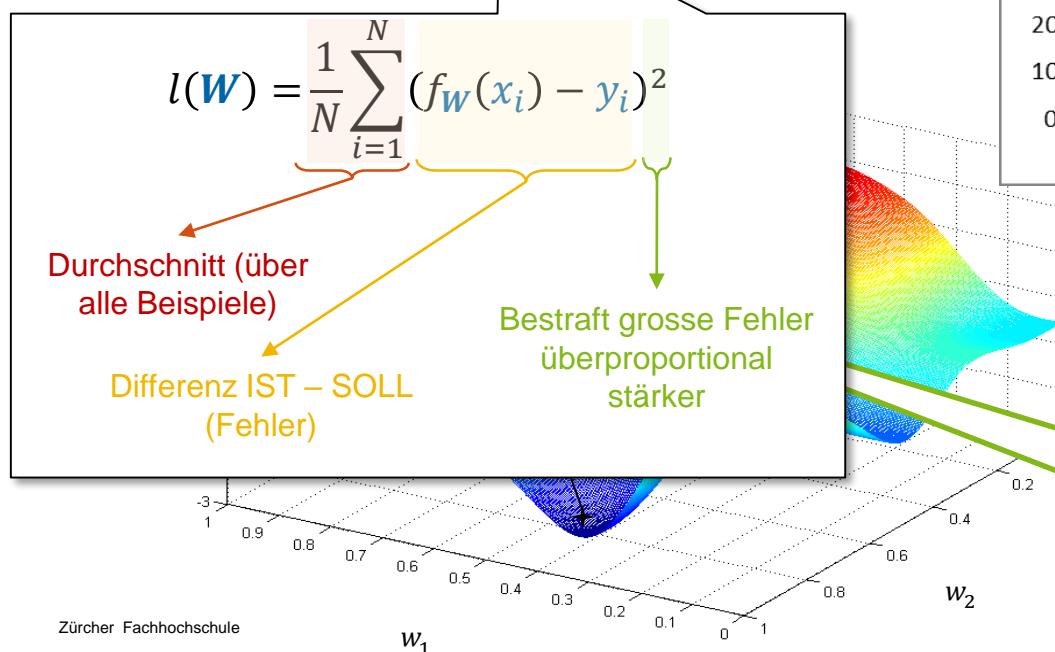
Making it actionable

- Publish a **Social Impact Statement**
- ...use above **principles as a guiding structure**
- ...revisit three times during development process: design stage, pre-launch, post-launch

Suche der Parameter einer Funktion?

Wahrscheinlichkeit [%] für bestimmtes Ergebnis

- Unser Neuronales Netz: $f_{\mathbf{W}}(\mathbf{x}) = \mathbf{y}$
mit Bild \mathbf{x} , echtem Resultat \mathbf{y} und Parametern \mathbf{W}
($\mathbf{W} = \{w_1, w_2, \dots\}$ anfangs zufällig gewählt)
- Fehlermass: $l(\mathbf{W}) = \frac{1}{N} \sum_{i=1}^N (f_{\mathbf{W}}(\mathbf{x}_i) - \mathbf{y}_i)^2$
Durchschnitt der quadratischen Abweichungen
über alle Bilder (Loss)



← Fehlerlandschaft

Methode: Anpassung der Gewichte
von f in Richtung der steilsten
Steigung (abwärts) von J

Was «sieht» das Neuronale Netz? Hierarchien komplexer werdender Merkmale

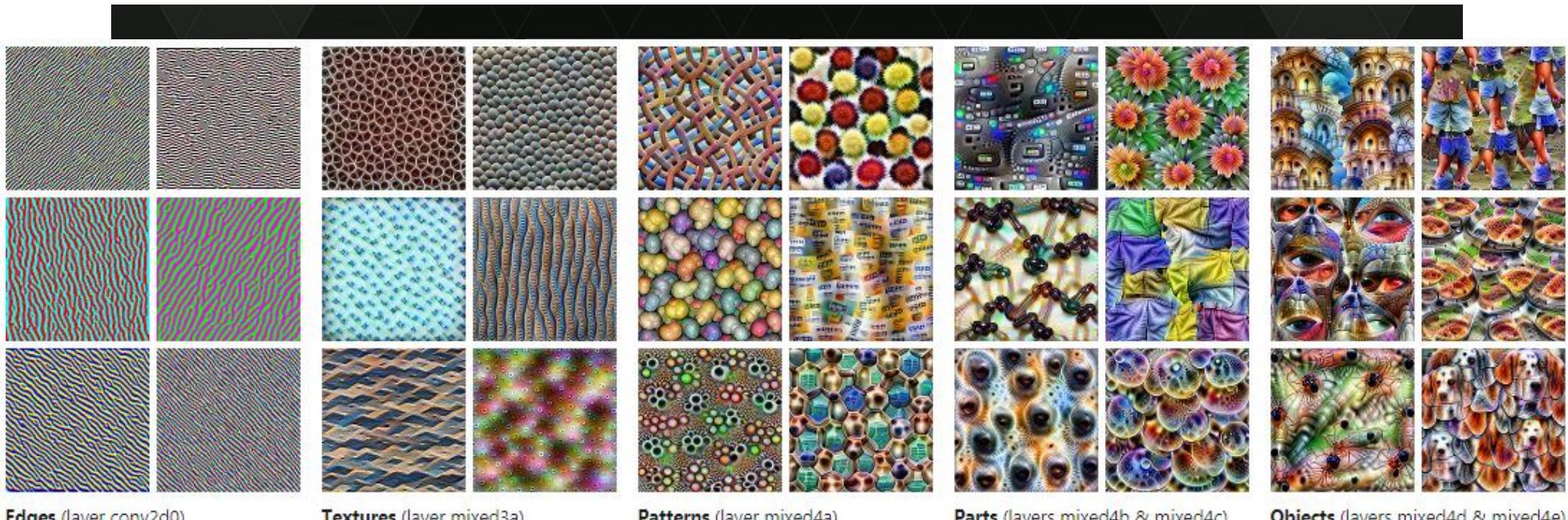


Image source: "Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks" ICML 2009 & Comm. ACM 2011.
Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Ng.

Quellen: <https://www.pinterest.com/explore/artificial-neural-network/>

Olah, et al., "Feature Visualization", Distill, 2017, <https://distill.pub/2017/feature-visualization/>.

Wie schlussfolgert die Maschine? «Debugging» für Einblicke in die vermeintliche «Black Box»

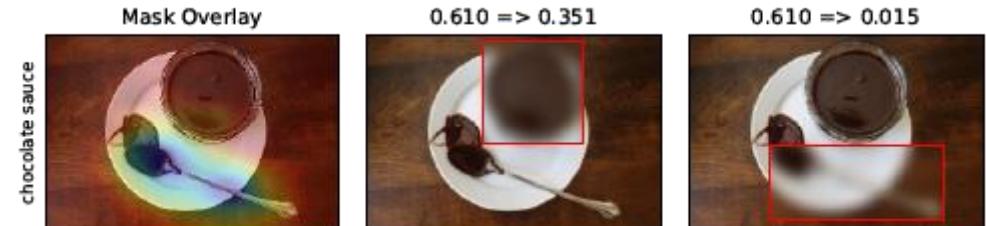
Verdeutlichen ein Problem:

- Adversarial Examples



Bieten eine Lösung:

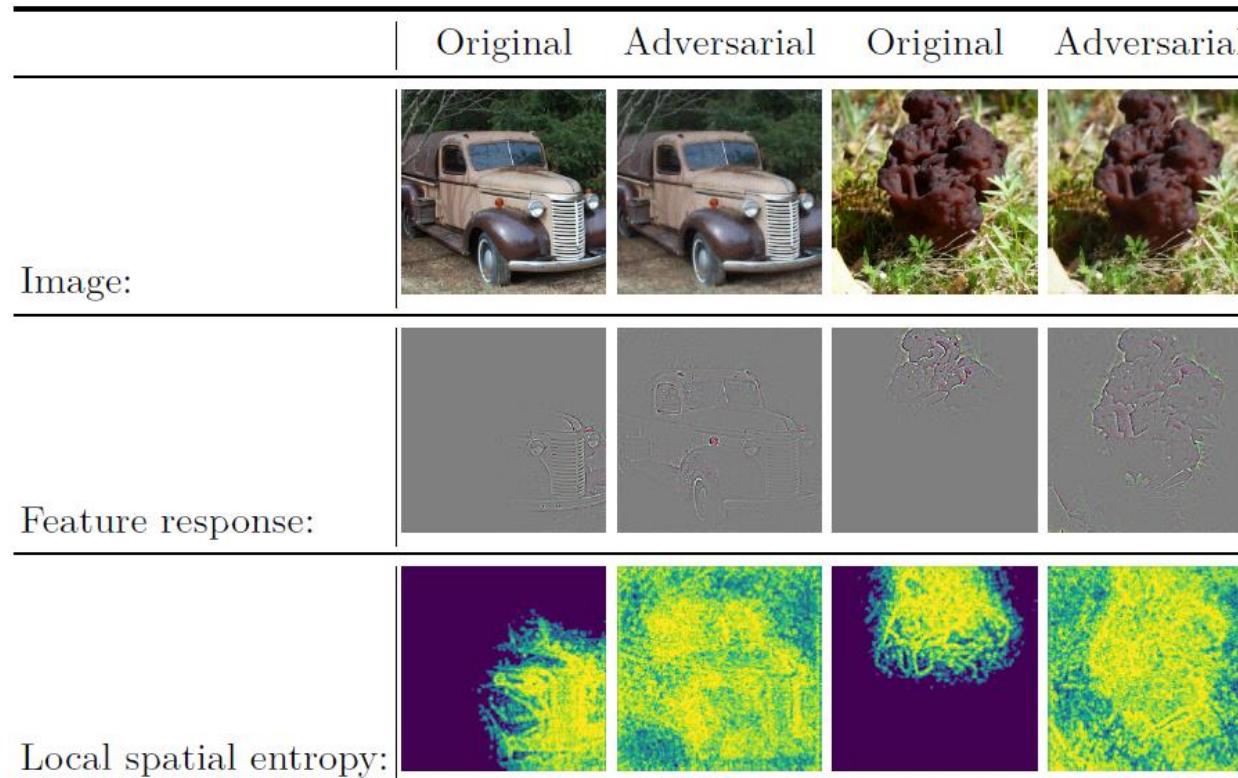
- Saliency Maps



Ruth C. Fong & Andrea Vedaldi, «Interpretable Explanations of Black Boxes by Meaningful Perturbation», 2017

Trace & detect adversarial attacks

...using average local spatial entropy of feature response maps



Amirian, Schwenker & Stadelmann (2018). «*Trace and Detect Adversarial Attacks on CNNs using Feature Response Maps*». ANNPR'2018.