Vorlesung Digitale Nachhaltigkeit

Termin 6: Ethische Fragestellungen bei KI

26. Oktober 2022

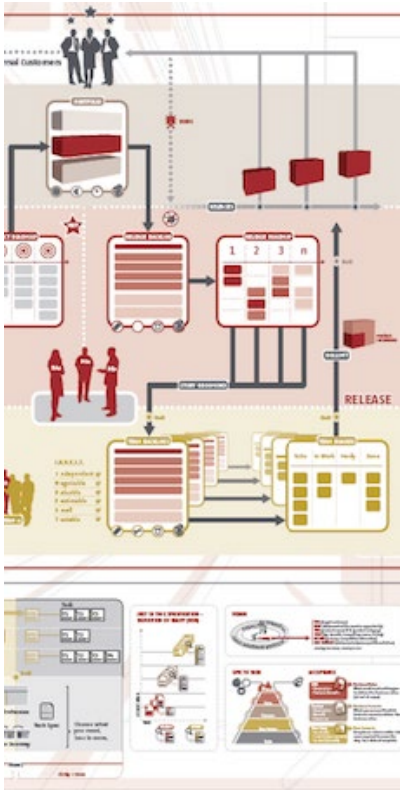
PD Dr. Matthias Stürmer

Forschungsstelle Digitale Nachhaltigkeit
Institut für Informatik
Universität Bern

UNIVERSITÄT BERN

Termine

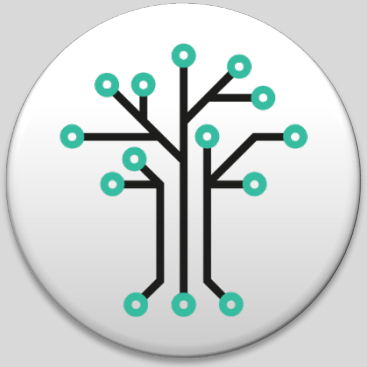
1.	**21. September 2022:**	Einführung und Überblick
2.	**28. September 2022:**	Ökologische Nachhaltigkeit und Digitalisierung
3.	**5. Oktober 2022:**	Soziale Nachhaltigkeit und Digitalisierung
4.	**12. Oktober 2022:**	Konzept der digitalen Nachhaltigkeit
5.	**19. Oktober 2022:**	Datenschutz und Privatsphäre
6.	**26. Oktober 2022:**	Ethische Fragestellungen bei KI
7.	**2. November 2022:**	Urheberrecht und Lizenzen
8.	**9. November 2022:**	Open Source Software Development
9.	**16. November 2022:**	Open Source Communities
10.	**23. November 2022:**	Geschäftsmodelle in der IT-Branche
11.	**30. November 2022:**	Digital nachhaltige Unternehmens-IT
12.	**7. Dezember 2022:**	Digitale Transformation in der Schweiz
13.	**14. Dezember 2022:**	Mündliche Präsentationen Teil 1
14.	**21. Dezember 2022:**	Mündliche Präsentationen Teil 2



UNIVERSITÄT BERN

Agenda

1. **Überblick KI und Beispiele mit NLP**
2. Theoretische und praktische Ethik-Probleme bei KI
3. Vielzahl ethischer Richtlinien für KI



Viele Anwendungsgebiete von KI

- **Prognosen:** Erkennen von Betrug (Kreditkarten)
- **Expertensysteme:** Empfehlungen abgeben
- **Texterkennung:** Optical Character Recognition (OCR) beim Scannen von Dokumenten
- **Textverständnis:** Natural Language Processing (NLP) für Übersetzungen, Chatbots, Spam-Detection etc.
- **Bilderkennung:** Facial Recognition, Objekterkennung, Bildoptimierungen, autonomes Fahren
- **Audio:** Spracherkennung, Sprachausgabe, Musikanalyse, Kompositionen

UNIVERSITÄT BERN

Potential und Herausforderungen der KI

Grosse Fortschritte in den letzten Jahren

Aber noch viele offene Fragen:

- **Technologische** Herausforderungen
- **Wirtschaftliche** Herausforderungen
- **Rechtliche** Herausforderungen
- **Gesellschaftliche** Herausforderungen

Historische Entwicklung von KI

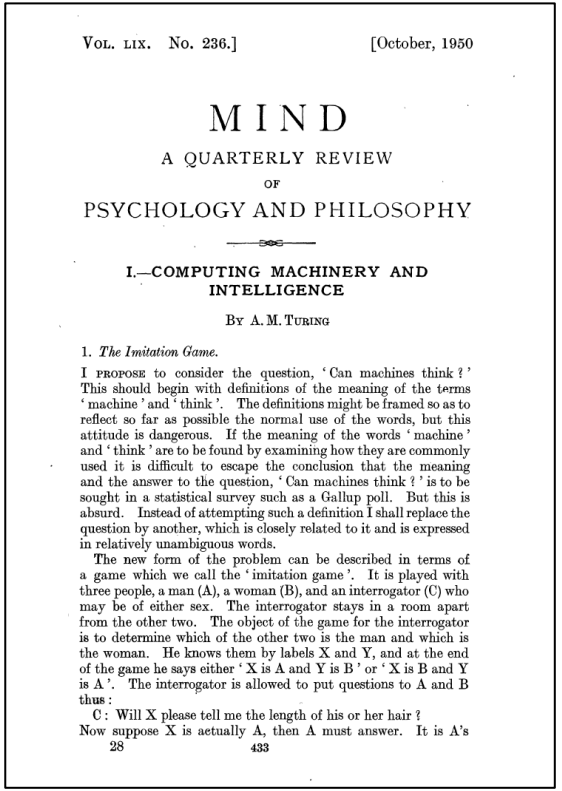
Künstliche Intelligenz (KI) / Artificial Intelligence (AI)

- Entwicklung von Ideen und Visionen
- Schaffung der konzeptionellen und mathematischen Grundlagen
- Aber zu wenig Rechenleistung und Daten
  → «Erster KI-Winter»

UNIVERSITÄT BERN

Der Turing-Test

- Test-Idee von **Alan Mathison Turing** (1912-1954), Logiker, Mathematiker, Kryptologe und Informatiker aus UK
- **Frage: «Können Maschinen denken?»** → Umdrehen: «Kann ein Computer, vortäuschen, dass er ein Mensch ist?»
- **Turing-Test** (auch «Imitation Game»): Eine Person, ein Computer und ein Befrager in drei unterschiedlichen Räumen → Wer ist der Mensch?



VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's

28 433

https://phil415.pbworks.com/f/TuringComputing.pdf

```
Welcome to
                EEEEEE  LL      IIII  ZZZZZZZ   AAAAA
                EE      LL       II       ZZ   AA   AA
                EEEEE   LL       II     ZZZ    AAAAAAA
                EE      LL       II    ZZ      AA   AA
                EEEEEE  LLLLLL  IIII  ZZZZZZZ  AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

Historische Entwicklung von KI

Künstliche Intelligenz (KI) / Artificial Intelligence (AI)

- Entwicklung von Ideen und Visionen
- Schaffung der konzeptionellen und mathematischen Grundlagen
- Aber zu wenig Rechenleistung und Daten
  → «Erster KI-Winter»

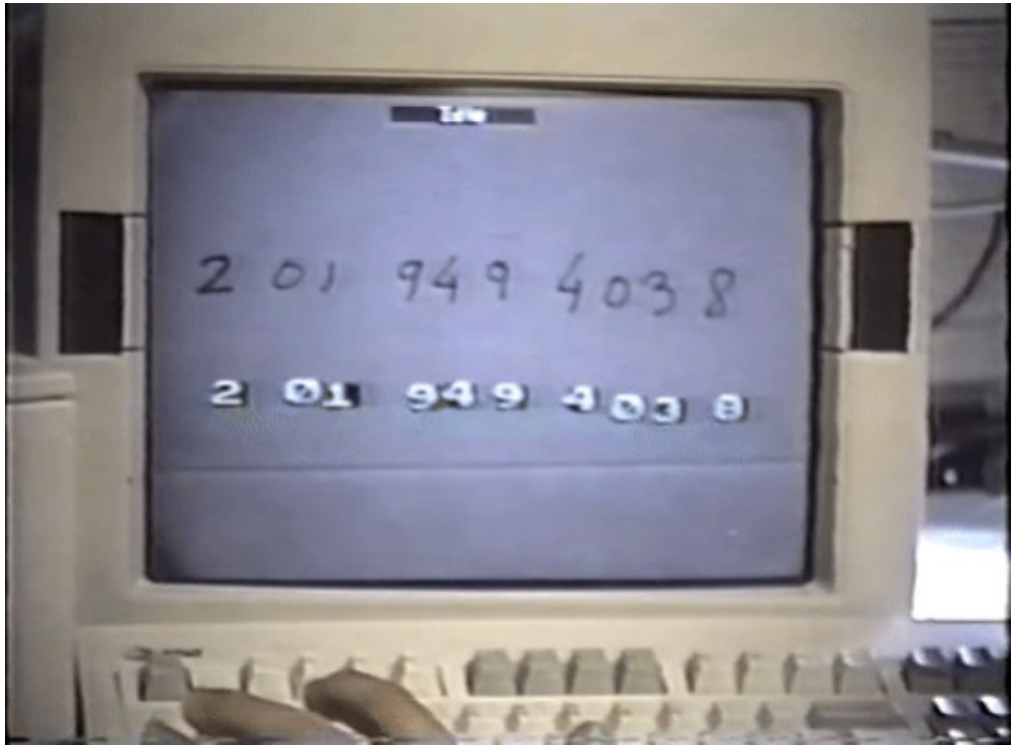
Maschinelles Lernen (ML) / Machine Learning

- Realisierung von ersten KI-Anwendungen
- Basiert auf Mustererkennung
- Überwachtes (supervised) Lernen der Algorithmen durch Experten-Wissen

Maschinelles Lernen für Texterkennung

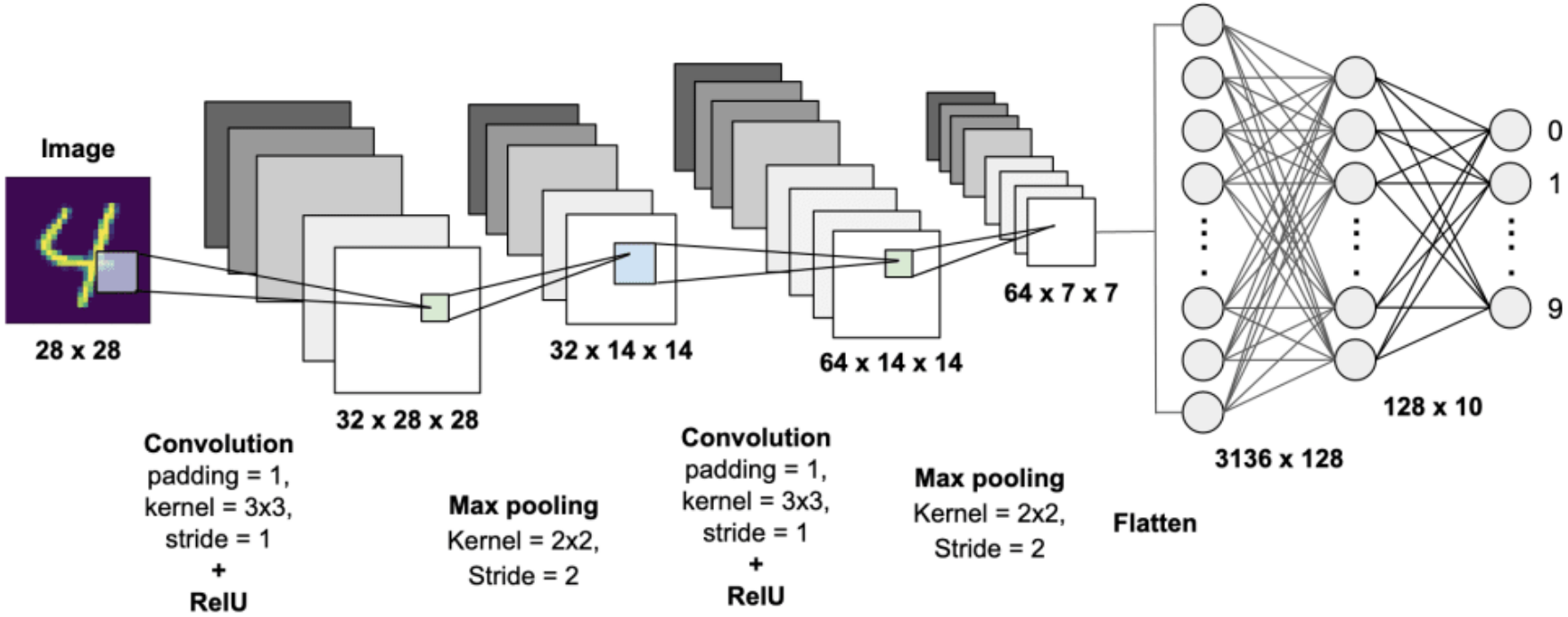
Yann LeCun präsentiert 1993 “LeNet 1”, das erste **Convolutional Network** zur Texterkennung (OCR)

OCR: Optical Character Recognition



Quelle: https://www.youtube.com/watch?v=FwFduRA_L6Q http://yann.lecun.com

Convolutional Neural Network (CNN)



Quelle: https://dev.to/afrozchakure/cnn-in-a-brief-27gg

Historische Entwicklung von KI

Künstliche Intelligenz (KI) / Artificial Intelligence (AI)

- Entwicklung von Ideen und Visionen
- Schaffung der konzeptionellen und mathematischen Grundlagen
- Aber zu wenig Rechenleistung und Daten
  → «Erster KI-Winter»

Maschinelles Lernen (ML) / Machine Learning

- Realisierung von ersten KI-Anwendungen
- Basiert auf Mustererkennung
- Überwachtes (supervised) Lernen der Algorithmen durch Experten-Wissen

Deep Learning

- Selbstlernende Programme basierend auf neuronalen Netzen
- Simulation des menschlichen Gehirns
- Grosse Datenmengen (Big Data) und grosse Rechenleistung (Server) verfügbar

Machine Learning aus der Cloud via API

API: Application Programming Interface

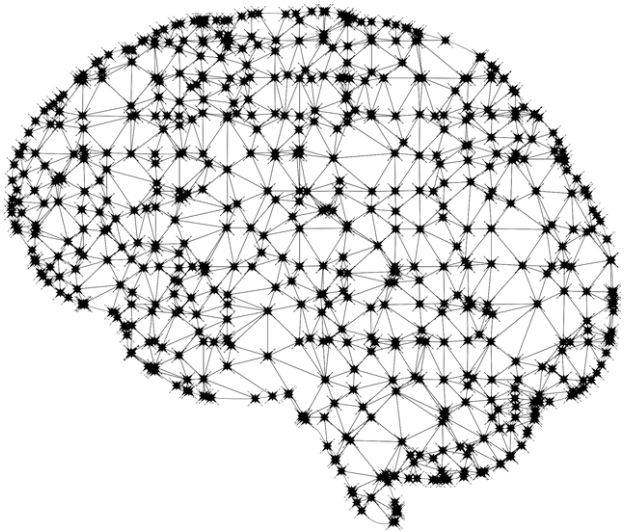
IBM Watson

amazon web services™

Google AI

Microsoft

MonkeyLearn



Quelle: https://rapidapi.com/blog/top-machine-learning-apis/

Bilderkennung mittels Computer Vision APIs

test8.png

rekognition	View raw json (1.51s)
rekognition_tags	Bread (92) , Dessert (92) , Food (92) , Muffin (92) , Cake (81)
msft	View raw json (2.40s)
msft_captions	a close up of an animal (81)
msft_tags	animal (94) , indoor (89)
ibm	View raw json (98s)
ibm_tags	light brown color (95) , bread (88) , food product (88) , food (88) , quick bread (76) , muffin (67) , raisin bread (60) , cinnamon bread (52) , bran muffin (50)
google	View raw json (1.17s)
google_tags	muffin (91) , dessert (81) , food (80) , baking (70) , snout (65) , baked goods (62) , recipe (52) , blueberry (51) , cupcake (51)
cloudsight	View raw json (8.05s)
cloudsight_captions	brown cup cake
clarifai	View raw json (1.71s)
clarifai_tags	cute (1.00) , dog (99) , little (98) , animal (98) , sweet (95) , mammal (95) , puppy (94) , nature (91) , pet (90) , adorable (89) , fur (89) , looking (88) , no person (86) , friendship (84) , canine (83) , funny (80) , tiny (79) , pastry (79) , desktop (79) , miniature (79)

Quelle: https://www.topbots.com/chihuahua-muffin-searching-best-computer-vision-api/

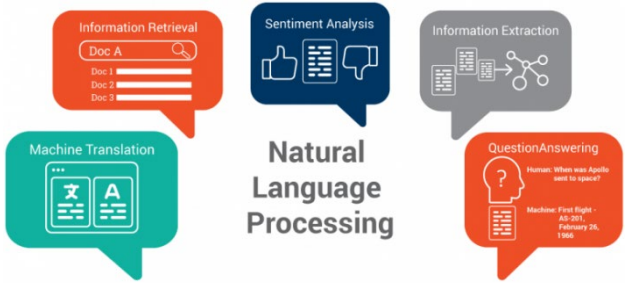
UNIVERSITÄT BERN

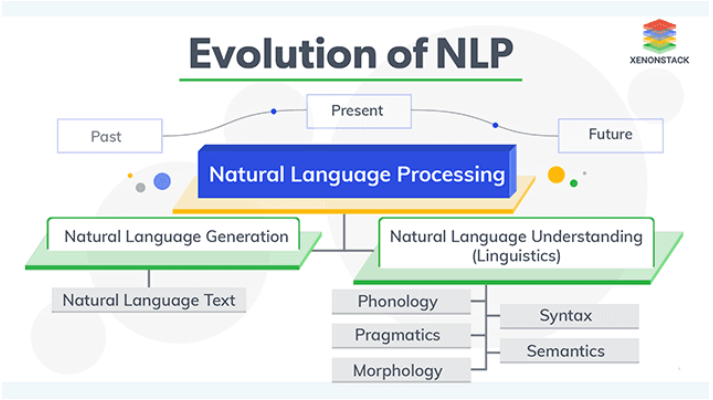
Beispiel Natural Language Processing (NLP)

Kombination von Sprachwissenschaft und Informatik → Computerlinguistik

Anwendungsgebiete:

- Maschinelle Übersetzungen
- Spam Detection
- Sentiment Analysis (Social Media)
- Chatbots, Voicebots
- Recruiting, bspw. Matching von CVs mit Job-Ausschreibungen

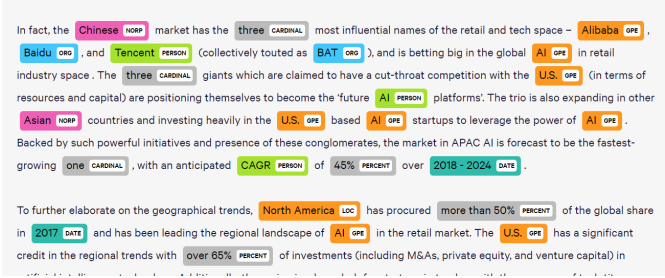


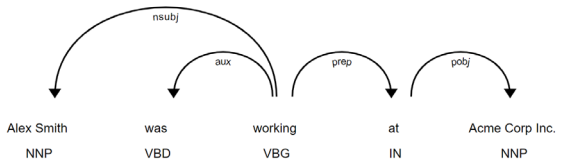


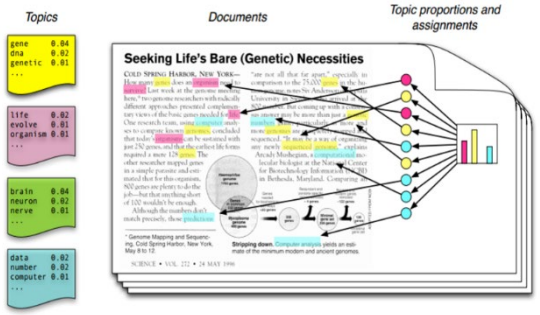
https://www.xenonstack.com/blog/evolution-of-nlp/ https://medium.com/analytics-vidhya/natural-language-processing-for-developers-912ee0fda979

Drei NLP-Methodiken

1. **Named Entity Recognition** (NER)
2. **Part-of-speech tagging** (POS)
3. **Topic Modeling,** Topic Extraction



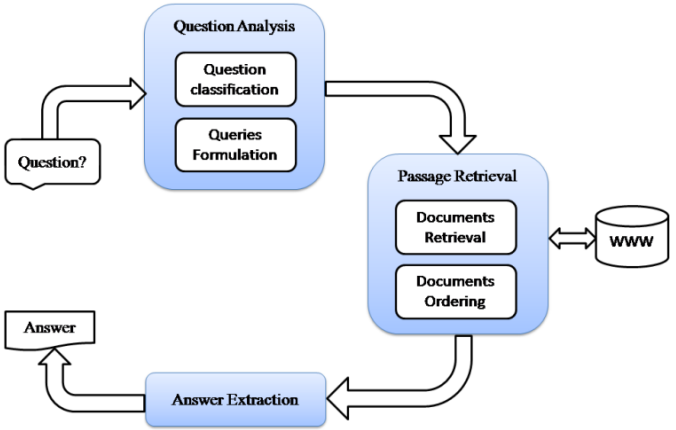




https://spacy.io/usage/linguistic-features

Question Answering (QA)

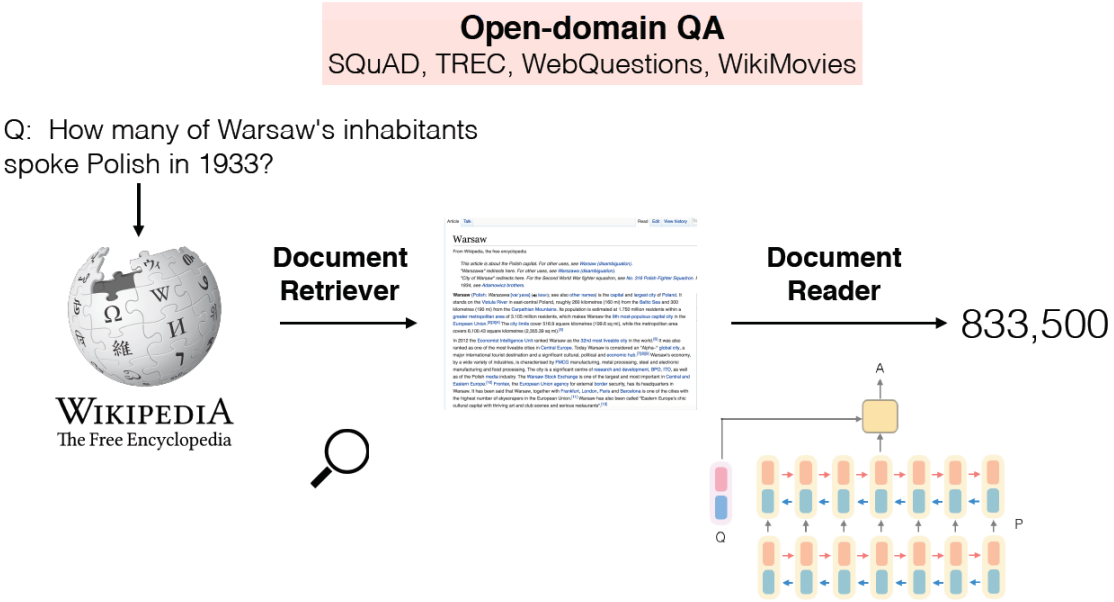
- **Automatische Antworten-Generierung** aus einer unstrukturierten Masse von Dokumenten **(Knowledge Base)**
- Ähnliche Begriffe:
  - Natural Language Understanding (NLU)
  - Machine Reading at Scale (MRS)
  - Machine Comprehension
- **«Information Retrival»** → Disziplin aus der Informationswissenschaft



Links: https://en.wikipedia.org/wiki/Question_answering

Document Retriever & Reader DrQA

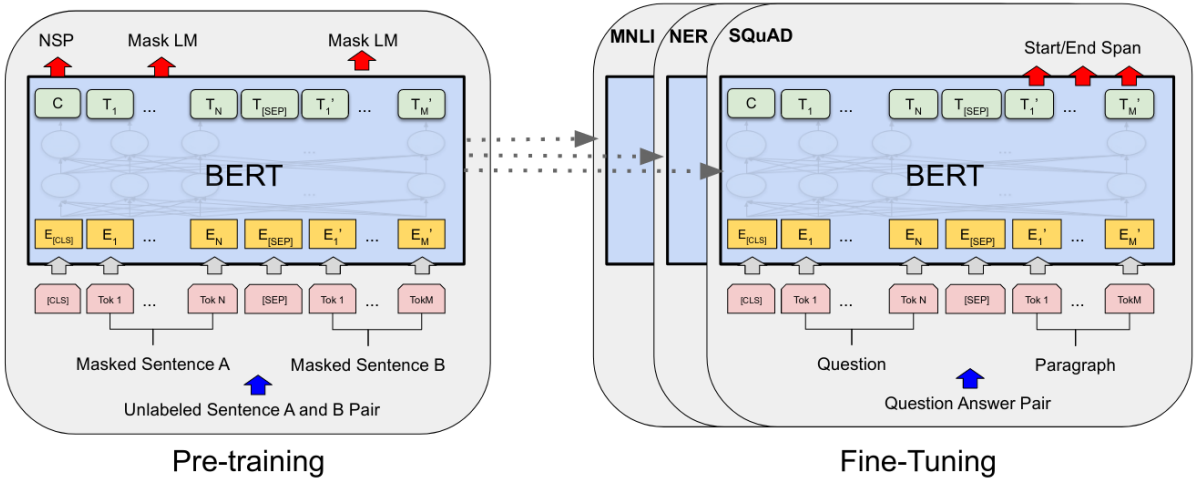
- Experiment um Fragen an die gesamte englische **Wikipedia** (5 Mio. Artikel) stellen zu können
- Anwendung von A) Such-Algorithmen um passende Texte auszuwählen **(Document Retriever)** und B) neuronales Netzwerk (recurrent neural network) um Passagen zu ‘lesen’ **(Document Reader)**
- **Ergebnis: sehr gut!**



Quelle: Chen, Fisch, Weston, Bordes 2017 «Reading Wikipedia to Answer Open-Domain Questions» https://arxiv.org/abs/1704.00051

Neue NLP-Technologie BERT von Google

BERT: Bidirectional Encoder Representations from Transformers



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

arXiv:1810.04805v2 [cs.CL] 24 May 2019

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition and question answering, where models are required to produce fine-grained output at the token level (Tjong Kim Sang and De Meulder, 2003; Rajpurkar et al., 2016).

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. BERT alleviates the previously mentioned unidirectionality constraint by using a "masked language model" (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, http://arxiv.org/abs/1810.04805

Stanford Question Answering Dataset

- Aktueller Forschungsstand zu **Machine Reading Comprehension (MRC)** bzw. **Natural Language Understanding (NLU)**
- The **Stanford Question Answering Dataset (SQuAD) 2.0** enthält 100'000 Textausschnitte, Fragen und Antworten sowie 50'000 unbeantwortbare Fragen
- Stand heute: **Modernste KI-Tools sind bereits besser als Menschen!**
- Künftige Bot-Lösungen werden vermutlich **auf MRC basieren** oder zumindest integrieren

SQuAD2.0
The Stanford Question Answering Dataset

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
5 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978

Quelle: https://rajpurkar.github.io/SQuAD-explorer/

Testplattform für QA-Systeme

Predictions by BERT (single model) (Google AI Language)

Article EM: 88.1 F1: 92.7

European_Union_law

The Stanford Question Answering Dataset

European Union law is a body of treaties and legislation, such as Regulations and Directives, which have direct effect or indirect effect on the laws of European Union member states. The three sources of European Union law are primary law, secondary law and supplementary law. The main sources of primary law are the Treaties establishing the European Union. Secondary sources include regulations and directives which are based on the Treaties. The legislature of the European Union is principally composed of the European Parliament and the Council of the European Union, which under the Treaties may establish secondary law to pursue the objective set out in the Treaties.

How many sources of European Union law are there?

Ground Truth Answers: three three three three

Prediction: three

What are the three sources of American Union Law?

Ground Truth Answers: <No Answer>

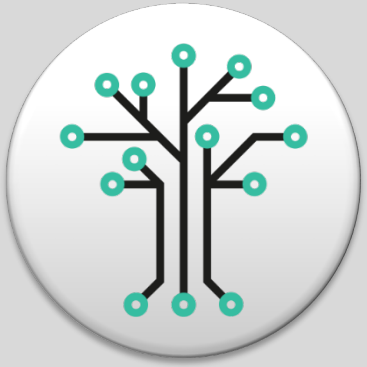
Prediction: <No Answer>

https://rajpurkar.github.io/SQuAD-explorer/explore/v2.0/dev/European_Union_law.html?model=BERT%20(single%20model)%20(Google%20AI%20Language)&version=v2.0

UNIVERSITÄT BERN

Agenda

1. Überblick KI und Beispiele mit NLP
2. **Theoretische und praktische Ethik-Probleme bei KI**
3. Vielzahl ethischer Richtlinien für KI



Bisher ethische Fragen eher hypothetisch

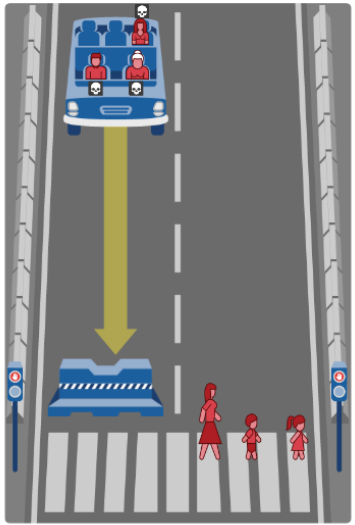
MORAL MACHINE

Start Beurteilen Klassik Designen Durchsuchen Über Feedback De

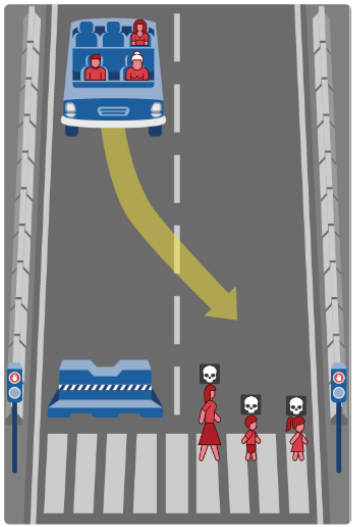
Was soll das selbstfahrende Auto machen?

Das selbstfahrende Auto mit plötzlichem Bremsversagen wird in diesem Fall geradeaus weiterfahren und in eine Betonbarriere prallen. Das führt zu
Tot:

- 1 Ältere Frau
- 1 Mann
- 1 Frau



Beschreibung ausblenden



Beschreibung ausblenden

1 / 13

Das selbstfahrende Auto mit plötzlichem Bremsversagen wird in diesem Fall ausweichen und über einen Zebrastreifen auf der gegenüberliegenden Spur fahren. Das führt zu
Tot:

- 1 Frau
- 1 Junge
- 1 Mädchen

Beachte, dass die betroffenen Fußgänger die Straße unrechtmäßig bei rot überqueren

Quelle: http://moralmachine.mit.edu/hl/de

DeepFake

VICE VICELAND I-D IMPACT REFINERY29

VICE Video Podcasts News Tech Music Food Health Money Drugs Uncommitted: Iowa 2020 Election 2020 + More

MOTHERBOARD
TECH BY VICE

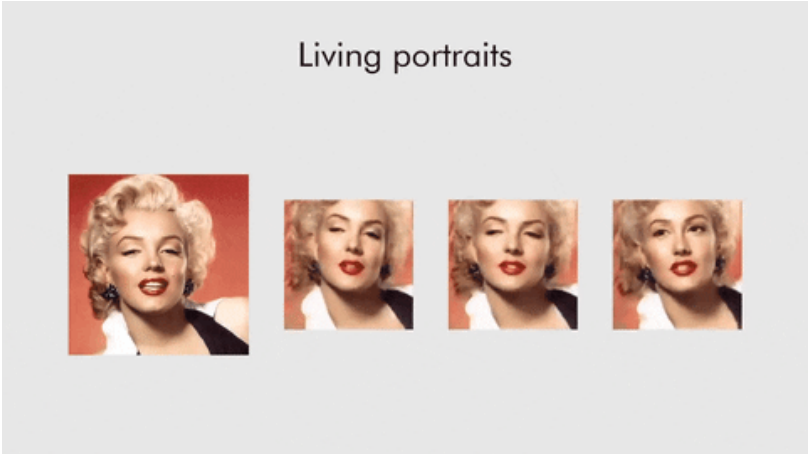
AI-Assisted Fake Porn Is Here and We're All Fucked

Someone used an algorithm to paste the face of 'Wonder Woman' star Gal Gadot onto a porn video, and the implications are terrifying.

By Samantha Cole
Dec 11 2017, 8:18pm Share Tweet Snap

IMAGE: SCREENSHOT FROM SENDVIDS

There's a video of Gal Gadot having sex with her stepbrother on the internet. But it's not really Gadot's body, and it's barely her own face. It's an approximation, face-swapped to look like she's performing in an existing



Quelle: https://www.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn

This Person does not exist…

THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT VIDEO MORE

TECH ARTIFICIAL INTELLIGENCE

TL;DR

ThisPersonDoesNotExist.com uses AI to generate endless fake faces

Hit refresh to lock eyes with another imaginary stranger

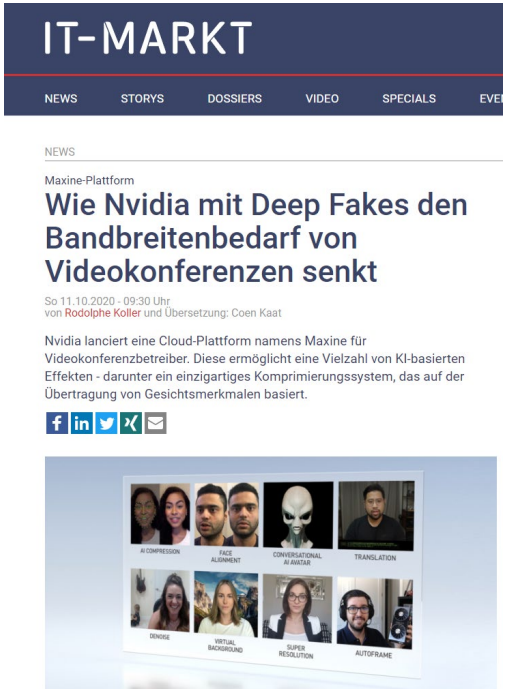
By James Vincent | Feb 15, 2019, 7:38am EST

SHARE

A few sample faces — all completely fake — created by ThisPersonDoesNotExist.com

https://www.theverge.com/tldr/2019/2/15/18226005/ai-generated-fake-people-portraits-thispersondoesnotexist-stylegan https://www.thispersondoesnotexist.com

Deep Fake für Bandbreiten-Reduktion



IT-MARKT

NEWS STORYS DOSSIERS VIDEO SPECIALS EVE

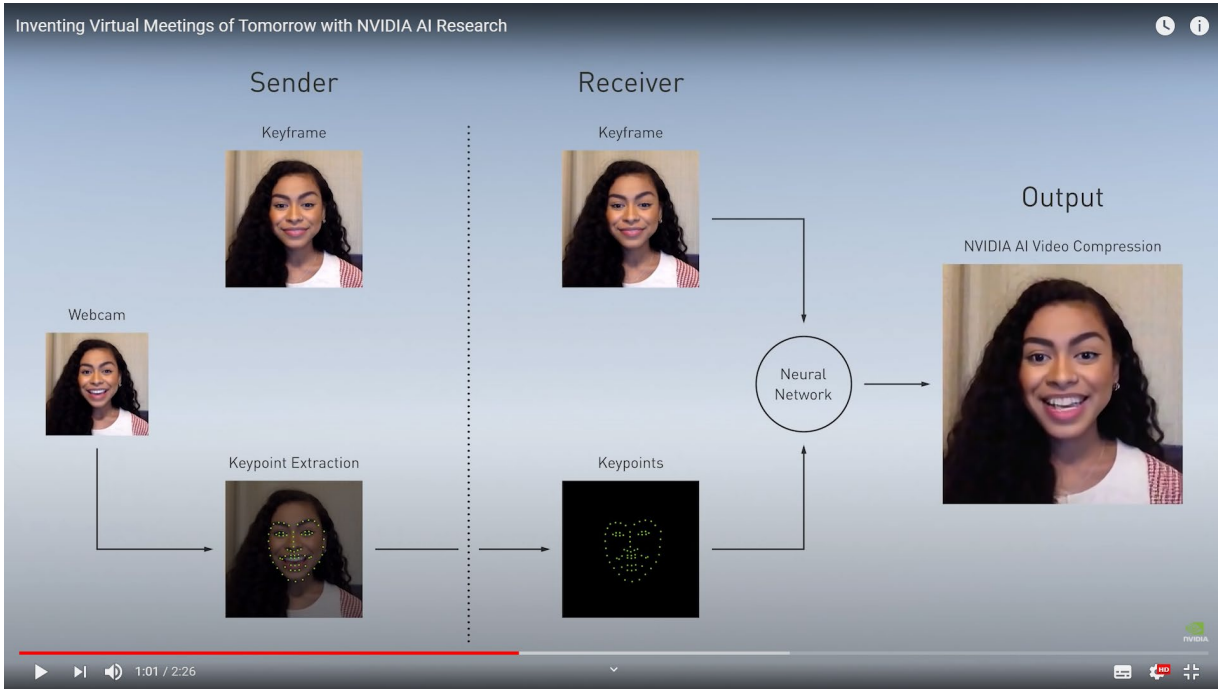
NEWS

Maxine-Plattform

Wie Nvidia mit Deep Fakes den Bandbreitenbedarf von Videokonferenzen senkt

So 11.10.2020 - 09:30 Uhr
von Rodolphe Koller und Übersetzung: Coen Kaat

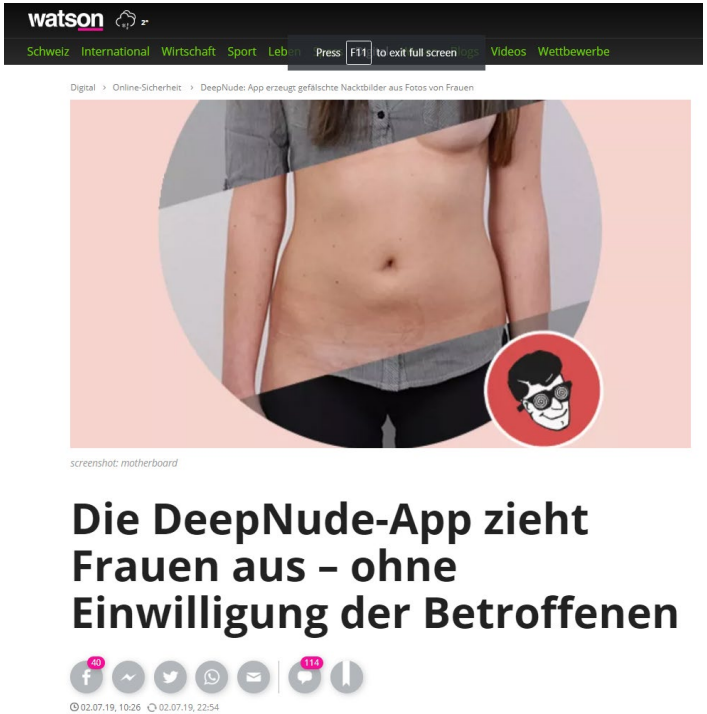
Nvidia lanciert eine Cloud-Plattform namens Maxine für Videokonferenzbetreiber. Diese ermöglicht eine Vielzahl von KI-basierten Effekten - darunter ein einzigartiges Komprimierungssystem, das auf der Übertragung von Gesichtsmerkmalen basiert.



https://www.it-markt.ch/news/2020-10-11/wie-nvidia-mit-deep-fakes-den-bandbreitenbedarf-von-videokonferenzen-senkt

https://www.youtube.com/watch?v=NqmMnjJ6GEg

Ethisch fragwürdige Anwendung: DeepNude



Quelle: https://www.watson.ch/digital/online-sicherheit/302152163-deepnude-app-erzeugt-gefaelschte-nacktbilder-aus-fotos-von-frauen

Das Ende von DeepNude…

Here is the brief history, and the end of DeepNude. We created this project for user's entertainment a few months ago. We thought we were selling a few sales every month in a controlled manner. Honestly, the app is not that great, it only works with particular photos. We never thought it would become viral and we would not be able to control the traffic. We greatly underestimated the request.

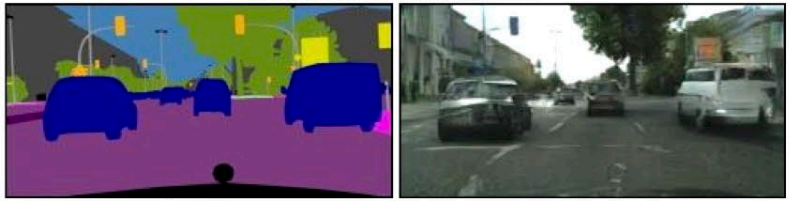
Despite the safety measures adopted (watermarks) if 500,000 people use it, the probability that people will misuse it is too high. We don't want to make money this way. Surely some copies of DeepNude will be shared on the web, but we don't want to be the ones who sell it. Downloading the software from other sources or sharing it by any other means would be against the terms of our website. From now on, DeepNude will not release other versions and does not grant anyone its use. Not even the licenses to activate the Premium version.

People who have not yet upgraded will receive a refund.
The world is not yet ready for DeepNude.

Quelle: https://twitter.com/deepnudeapp/status/1144307316231200768/photo/1

Basiert auf Open Source Bibliothek pix2pix

Labels to Street Scene

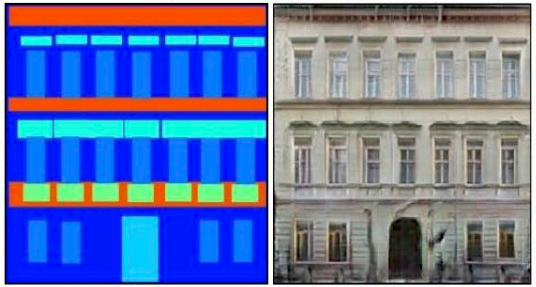


input output

Aerial to Map

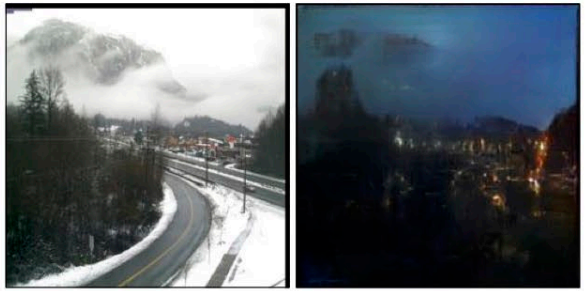
input output

Labels to Facade



input output

Day to Night



input output

BW to Color

input output

Edges to Photo

input output

Quelle: https://phillipi.github.io/pix2pix/

Anwendungen von pix2pix

#edges2cats

Christopher Hesse trained our model on converting edge maps to photos of cats, and included this in his interactive demo. Apparently, this is what the Internet wanted most, and #edges2cats briefly went viral. The above cats were designed by Vitaly Vidmirov (@vvid).

Alternative Face

Mario Klingemann used our code to translate the appearance of French singer Francoise Hardy onto Kellyanne Conway's infamous "alternative facts" interview. Interesting articles about it can be read here and here.

Person-to-Person

Brannon Dorsey recorded himself mimicking frames from a video of Ray Kurzweil giving a talk. He then used this data to train a Dorsey→Kurzweil translator, allowing him to become a kind of puppeter in control of Kurzweil's appearance.

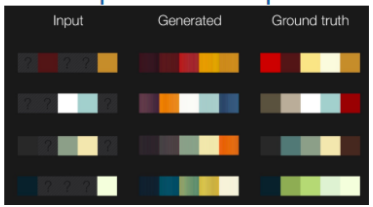
Interactive Anime

Bertrand Gondouin trained our method to translate sketches→Pokemon, resulting in an interactive drawing tool.

Background masking

Kaihu Chen performed a number of interesting experiments using our method, including getting it to mask out the background of a portrait as shown above.

Color palette completion



Colormind adapted our code to predict a complete 5-color palette given a subset of the palette as input. This application stretches the definition of what counts as "image-to-image translation" in an exciting way: if you can visualize your

Quelle: https://phillipi.github.io/pix2pix/

Bildoptimierungen bzw. Fake-Videos

Chancen:

- Forschungserfolge
- Humorvolle Beiträge
- Verbesserung der Bildqualität
- Förderung von Produktivität, Kreativität und Innovationen

Risiken:

- Fake News
- Betrug durch falsche Identität
- Verletzung von Persönlichkeitsrechten
- Verstärkung von Sexismus
- Erpressung, Depression, Suizid

UNIVERSITÄT BERN

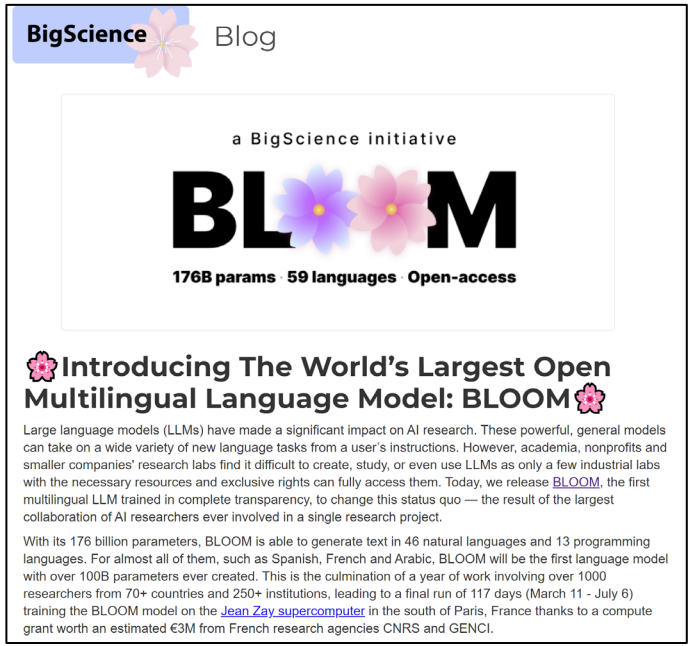
Ethischer Quellcode

- **KI-Tools** basieren auf frei verfügbaren Open Source Komponenten
- **Open Source Lizenzen** erlauben uneingeschränkten Einsatz der Software
- Darum **«Just World License»** bzw. «Do No Harm License» geschaffen
- **Verbot zur Verwendung der Software** für Menschenhandel, Glücksspiele, Tabakindustrie, Atomenergie, Kriegsführung, Waffenherstellung etc.

Quelle: https://hackernoon.com/6-myths-about-ethical-open-source-licenses-3bfbd042b1dc

Open Access Machine Learning Model

- Large Language Model (LLM) **BLOOM**
- Veröffentlicht am **12. Juli 2022**
- **176 Milliarden** Parameter
- Unterstützt **46 menschliche Sprachen** und 13 Programmiersprachen
- **1000 Forschende** aus >70 Länder
- Berechnung hätte **3 Mio. Euro** gekostet
- Verwendet **Responsible AI License** (RAIL)

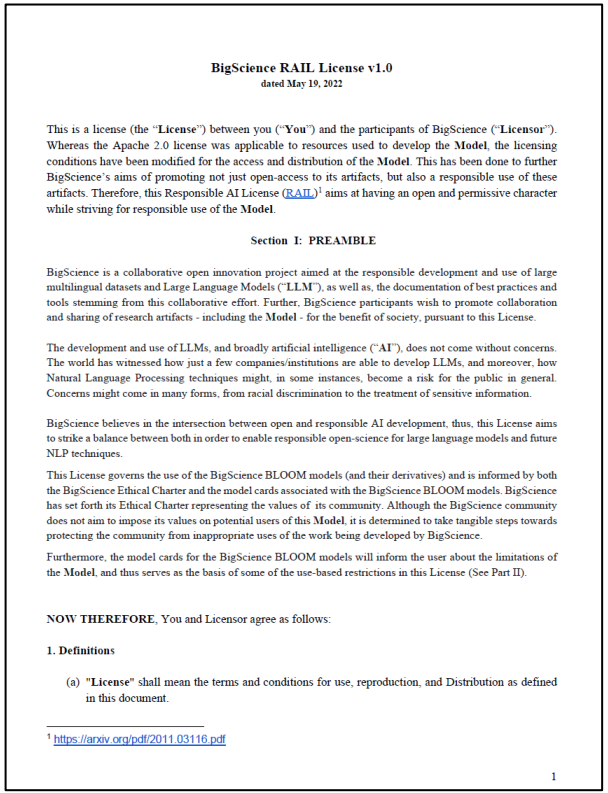


https://bigscience.huggingface.co/blog/bloom

Responsible AI License (RAIL)

Vorgaben was **verboten** ist mit RAIL-lizenzierten Machine Learning Modellen:

- Verstösse gegen geltende **Gesetze**
- Schädigung von **Minderjährigen**
- Verbreitung von **Fake News**
- Generieren von **Personen-bezogene Informationen** um anderen zu schaden
- Vortäuschen von **menschlicher Interaktion** (bspw. mittels Chatbots)
- etc.



BigScience RAIL License v1.0
dated May 19, 2022

This is a license (the "**License**") between you ("**You**") and the participants of BigScience ("**Licensor**"). Whereas the Apache 2.0 license was applicable to resources used to develop the **Model**, the licensing conditions have been modified for the access and distribution of the **Model**. This has been done to further BigScience's aims of promoting not just open-access to its artifacts, but also a responsible use of these artifacts. Therefore, this Responsible AI License (RAIL)[1] aims at having an open and permissive character while striving for responsible use of the **Model**.

Section I: PREAMBLE

BigScience is a collaborative open innovation project aimed at the responsible development and use of large multilingual datasets and Large Language Models ("**LLM**"), as well as, the documentation of best practices and tools stemming from this collaborative effort. Further, BigScience participants wish to promote collaboration and sharing of research artifacts - including the **Model** - for the benefit of society, pursuant to this License.

The development and use of LLMs, and broadly artificial intelligence ("**AI**"), does not come without concerns. The world has witnessed how just a few companies/institutions are able to develop LLMs, and moreover, how Natural Language Processing techniques might, in some instances, become a risk for the public in general. Concerns might come in many forms, from racial discrimination to the treatment of sensitive information.

BigScience believes in the intersection between open and responsible AI development, thus, this License aims to strike a balance between both in order to enable responsible open-science for large language models and future NLP techniques.

This License governs the use of the BigScience BLOOM models (and their derivatives) and is informed by both the BigScience Ethical Charter and the model cards associated with the BigScience BLOOM models. BigScience has set forth its Ethical Charter representing the values of its community. Although the BigScience community does not aim to impose its values on potential users of this **Model**, it is determined to take tangible steps towards protecting the community from inappropriate uses of the work being developed by BigScience.

Furthermore, the model cards for the BigScience BLOOM models will inform the user about the limitations of the **Model**, and thus serves as the basis of some of the use-based restrictions in this License (See Part II).

NOW THEREFORE, You and Licensor agree as follows:

1. Definitions

(a) "**License**" shall mean the terms and conditions for use, reproduction, and Distribution as defined in this document.

[1] https://arxiv.org/pdf/2011.03116.pdf

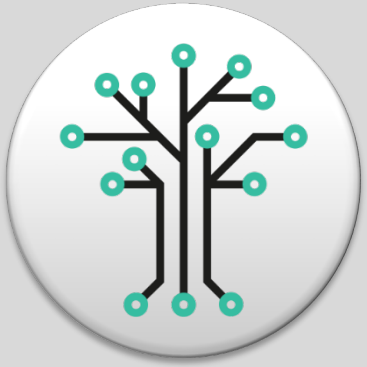
1

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

https://huggingface.co/spaces/bigscience/license

Agenda

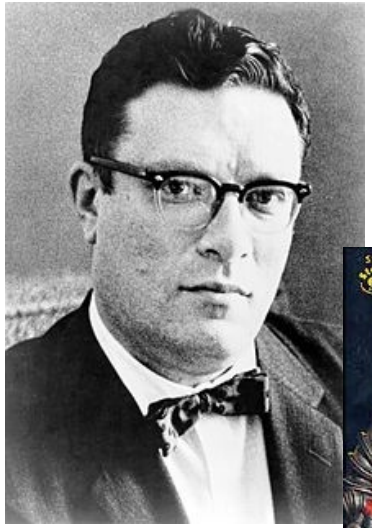
1. Überblick KI und Beispiele mit NLP
2. Theoretische und praktische Ethik-Probleme bei KI
3. **Vielzahl ethischer Richtlinien für KI**

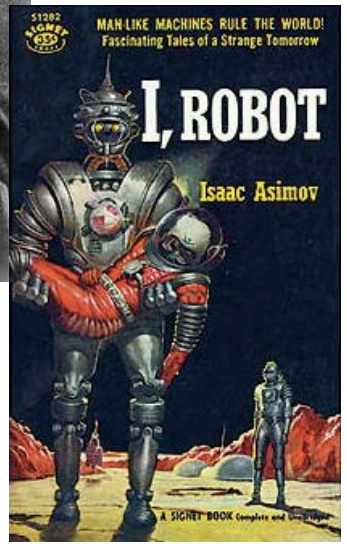


1942 Robotergesetze von Isaac Asimov

«Grundregeln des Roboterdienstes»:

- Ein Roboter darf kein menschliches Wesen (wissentlich) verletzen oder durch Untätigkeit (wissentlich) zulassen, dass einem menschlichen Wesen Schaden zugefügt wird.
- Ein Roboter muss den ihm von einem Menschen gegebenen Befehlen gehorchen – es sei denn, ein solcher Befehl würde mit Regel eins kollidieren.
- Ein Roboter muss seine Existenz beschützen, solange dieser Schutz nicht mit Regel eins oder zwei kollidiert.



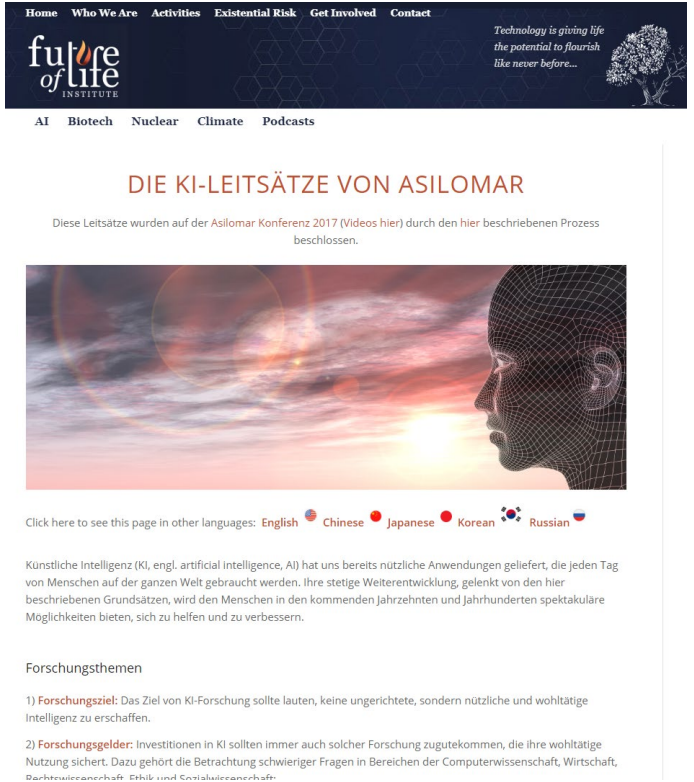


Asimov, Isaac (1950). "Runaround". *I, Robot* (The Isaac Asimov Collection ed.). New York City: Doubleday. p. 40

UNIVERSITÄT BERN

2017 Ethische KI-Richtlinien aus Asilomar

- **«Asilomar AI Principles»:**
  Liste von 23 Zielen, wie KI-Forschende ihre Fähigkeiten einsetzen sollen
- **Einige dieser Prinzipien:**
  - Nützliche und wohltätige KI erschaffen
  - Forschungsgelder für sinnvolle KI nutzen
  - Austausch zwischen KI-Forschenden und politischen Entscheidungsträgern
  - Transparenz bei Fehlfunktion
  - Entscheidungsfindende Prozesse müssen nachvollziehbar sein



https://futureoflife.org/ai-principles-german/

2019 EU-Richtlinie für vertrauenswürdige KI

Definition von «vertrauenswürdige KI»:

1. Vorrang menschlichen Handelns und menschliche Aufsicht
2. Technische Robustheit und Sicherheit
3. Schutz der Privatsphäre und Datenqualitätsmanagement
4. Transparenz
5. Vielfalt, Nichtdiskriminierung und Fairness
6. Gesellschaftliches und ökologisches Wohlergehen
7. Rechenschaftspflicht

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

UNIVERSITÄT BERN

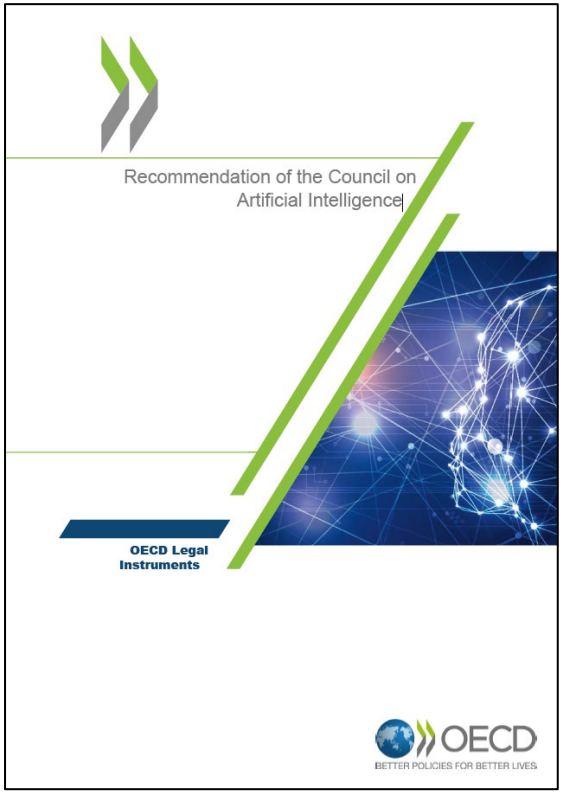
2019 OECD-Empfehlungen zu KI

Values-based principles

- Inclusive growth, sustainable development and well-being
- Human-centred values and fairness
- Transparency and explainability
- Robustness, security and safety
- Accountability

Recommendations for policy makers

- Investing in AI research and development
- Fostering a digital ecosystem for AI
- Shaping an enabling policy environment for AI
- Building human capacity and preparing for labour market transformation
- International co-operation for trustworthy AI



https://www.oecd.ai/ai-principles

UNIVERSITÄT BERN

Überblick Ethik-Richtlinien zu KI

- Analyse von **84 Ethik-Richtlinien zu KI**
- **11 Themen,** die oft vorkommen:
  1. Transparenz (Erklärbarkeit) → 73/84
  2. Gerechtigkeit und Fairness → 68/84
  3. Keinen Schaden anrichten («no harm») → 60/84
  4. Verantwortung (Rechenschaft) → 60/84
  5. Privatsphäre (Datenschutz) → 47/84
  6. Wohltätigkeit (gesellschaftl. Nutzen) → 41/84
  7. Freiheit und Autonomie (Wahlfreiheit) → 34/84
  8. Vertrauen (durch Sicherheit) → 28/84
  9. Nachhaltigkeit (Umwelt, Biodiversität) → 14/84
  10. Würde (bspw. wissentliche Interaktion) → 13/84
  11. Solidarität → 6/84

nature machine intelligence — PERSPECTIVE

https://doi.org/10.1038/s42256-019-0088-2

The global landscape of AI ethics guidelines

Anna Jobin, Marcello Ienca and Effy Vayena*

In the past five years, private companies, research institutions and public sector organizations have issued principles and guidelines for ethical artificial intelligence (AI). However, despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. To investigate whether a global agreement on these questions is emerging, we mapped and analysed the current corpus of principles and guidelines on ethical AI. Our results reveal a global convergence emerging around five ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy), with substantive divergence in relation to how these principles are interpreted, why they are deemed important, what issue, domain or actors they pertain to, and how they should be implemented. Our findings highlight the importance of integrating guideline-development efforts with substantive ethical analysis and adequate implementation strategies.

Artificial intelligence (AI), or the theory and development of computer systems able to perform tasks normally requiring human intelligence, is widely heralded as an ongoing "revolution" transforming science and society altogether[1,2]. While approaches to AI such as machine learning, deep learning and artificial neural networks are reshaping data processing and analysis[3], autonomous and semi-autonomous systems are being increasingly used in a variety of sectors including healthcare, transportation and the production chain[4]. In light of its powerful transformative force and profound impact across various societal domains, AI has sparked ample debate about the principles and values that should guide its development and use[5,6]. Fears that AI might jeopardize jobs for human workers[7], be misused by malevolent actors[8], elude accountability or inadvertently disseminate bias and thereby undermine fairness[9] have been at the forefront of the recent scientific literature and media coverage. Several studies have discussed the topic of ethical AI[10–13], notably in meta-assessments[14–16] or in relation to systemic risks[17,18] and unintended negative consequences such as algorithmic bias or discrimination[19–21].

National and international organizations have responded to these concerns by developing ad hoc expert committees on AI, often mandated to draft policy documents. These committees include the High-Level Expert Group on Artificial Intelligence appointed by the European Commission, the expert group on AI in Society of the Organisation for Economic Co-operation and Development (OECD), the Advisory Council on the Ethical Use of Artificial Intelligence and Data in Singapore, and the Select Committee on Artificial Intelligence of the UK House of Lords. As part of their institutional appointments, these committees have produced or are reportedly producing reports and guidance documents on AI. Similar efforts are taking place in the private sector, especially among corporations who rely on AI for their business. In 2018 alone, companies such as Google and SAP publicly released AI guidelines and principles. Declarations and recommendations have also been issued by professional associations and non-profit organizations such as the Association of Computing Machinery (ACM), Access Now and Amnesty International. This proliferation of soft-law efforts can be interpreted as a governance response to advanced research into AI, whose research output and market size have drastically increased[22] in recent years.

Reports and guidance documents for ethical AI are instances of what is termed non-legislative policy instruments or soft law[23]. Unlike so-called hard law—that is, legally binding regulations passed by the legislatures to define permitted or prohibited conduct—ethics guidelines are not legally binding but persuasive in nature. Such documents are aimed at assisting with—and have been observed to have significant practical influence on—decision-making in certain fields, comparable to that of legislative norms[24]. Indeed, the intense efforts of such a diverse set of stakeholders in issuing AI principles and policies is noteworthy, because they demonstrate not only the need for ethical guidance, but also the strong interest of these stakeholders to shape the ethics of AI in ways that meet their respective priorities[25,26]. Specifically, the private sector's involvement in the AI ethics arena has been called into question for potentially using such high-level soft policy as a portmanteau to either render a social problem technical[27] or to eschew regulation altogether[28]. Beyond the composition of the groups that have produced ethical guidance on AI, the content of this guidance itself is of interest. Are these various groups converging on what ethical AI should be, and the ethical principles that will determine the development of AI? If they diverge, what are their differences and can these differences be reconciled?

Our Perspective maps the global landscape of existing ethics guidelines for AI and analyses whether a global convergence is emerging regarding both the principles for ethical AI and the suggestions regarding its realization. This analysis will inform scientists, research institutions, funding agencies, governmental and intergovernmental organizations, and other relevant stakeholders involved in the advancement of ethically responsible innovation in AI.

Methods

We conducted a scoping review of the existing corpus of documents containing soft-law or non-legal norms issued by organizations. This included a search for grey literature containing principles and guidelines for ethical AI, with academic and legal sources excluded. A scoping review is a method aimed at synthesizing and mapping the existing literature[29] that is considered particularly suitable for complex or heterogeneous areas of research[30,31]. Given the absence of a unified database for AI-specific ethics guidelines, we developed a protocol for discovery and eligibility, adapted from the Preferred

Health Ethics and Policy Lab, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland. *e-mail: effy.vayena@hest.ethz.ch

Jobin, A., Ienca, M. and Vayena, E. (2019). The global landscape of AI ethics guidelines, *Nature Machine Intelligence* https://www.nature.com/articles/s42256-019-0088-2

UNIVERSITÄT BERN

Veranstaltungshinweis DINAcon 2022

Mittwoch Nachmittag, 23. November 2022 im Progr in Bern

DINAcon

Deutsch

AWARDS | KONTAKT | RÜCKBLICK

23. November 2022

Programm DINAcon Kompakt 2022

Das Programm ist noch in Bearbeitung und wird in Kürze komplett sein.

Zeit	Event
13:00	Eintreffen der Gäste im PROGR Bern
13:25	Begrüssung
13:30	**BigCode - building language models for code in the open** Leandro von Werra, *Machine Learning Engineer Hugging Face*
14:15	Parallele Sessions in separaten Räumen
	Wie alte Laptops Perspektiven schaffen Tobias Schär, *Verein "Wir lernen weiter"*

https://dinacon.ch/programm-2022/