KI in der Praxis: Chancen und Gefahren

Prof. Dr. **Dominik Herrmann**

LSt Privatsphäre und Sicherheit
in Informationssystemen
Universität Bamberg

Slides: https://dhgo.to/winfor-ki

https://deepfriedbits.com/2019/10/31/go-home-roomba-youre-drunk/

CNN BUSINESS Markets Tech Media Success Video

Control AI now or brace for nightmare future, experts warn

by Sherisse Pham @Sherisse

February 21, 2018: 6:58 AM ET

TIME SUBSCRIBE

ROBOTICS

5 Very Smart People Who Think Artificial Intelligence Could Bring the Apocalypse

FUTURE SOCIETY

Kurzweil Claims That the Singularity Will Happen by 2045

Get ready for humanity 2.0.

Eine Frage der

Moral

22.03.2018

Uber

Erster Todesfall mit Roboterauto heizt Sicherheits-Debatte an

Was soll ein selbst-fahrendes Auto tun, wenn es nicht mehr rechtzeitig bremsen kann?

Wie soll sich das Auto verhalten?

Wie ein Mensch

zufällig

Wie eine Maschine

minimaler Schaden

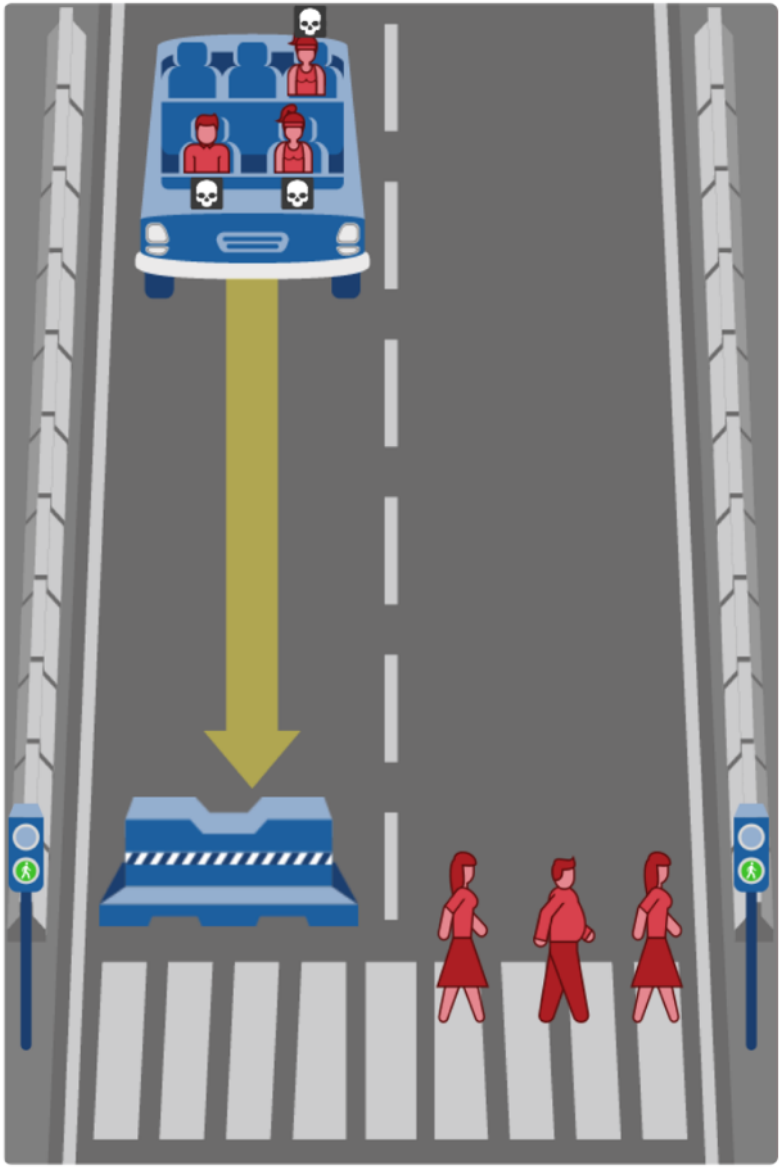
Wer legt „minimal“ fest?
Entwickler? Experten?
Oder der Passagier?

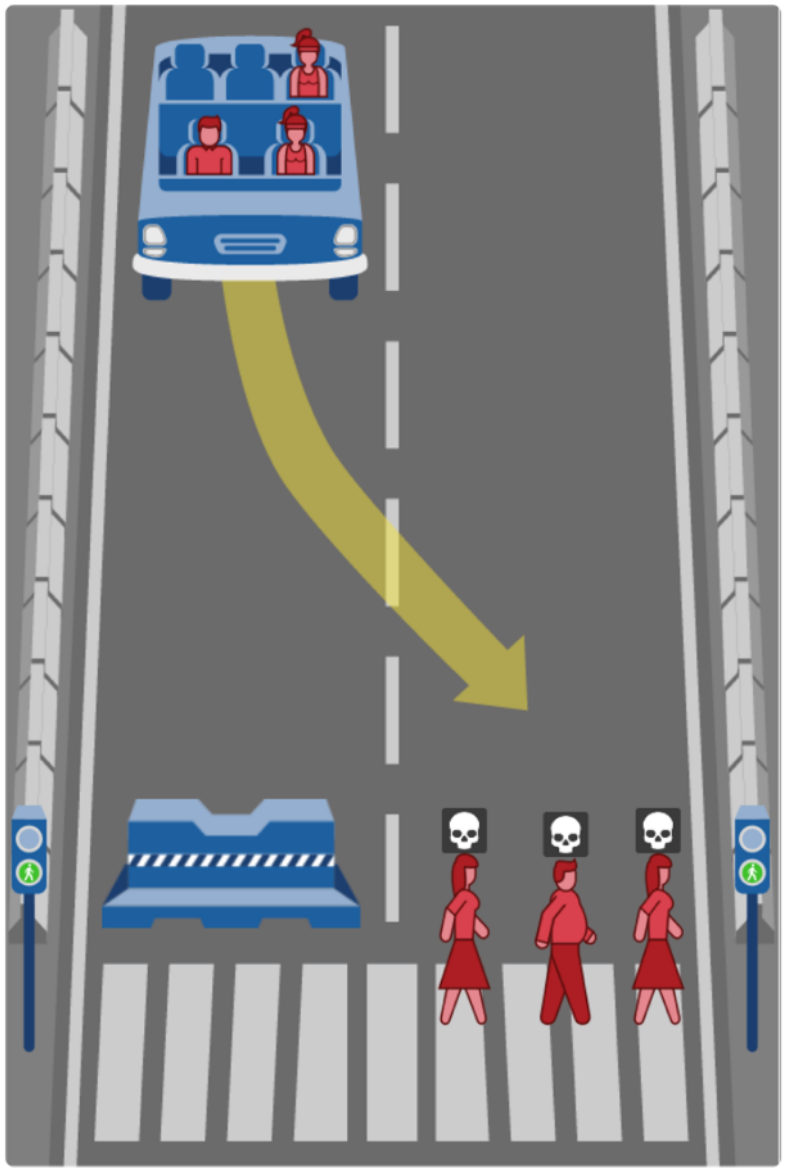
Eine Situation in der Praxis

ohne Helm	**mit Helm**
stirbt wahrscheinlich (und ist ein VIP!)	überlebt vielleicht

http://www.theguardian.com/environment/green-living-blog/2010/apr/07/david-cameron-cycles-without-helmet

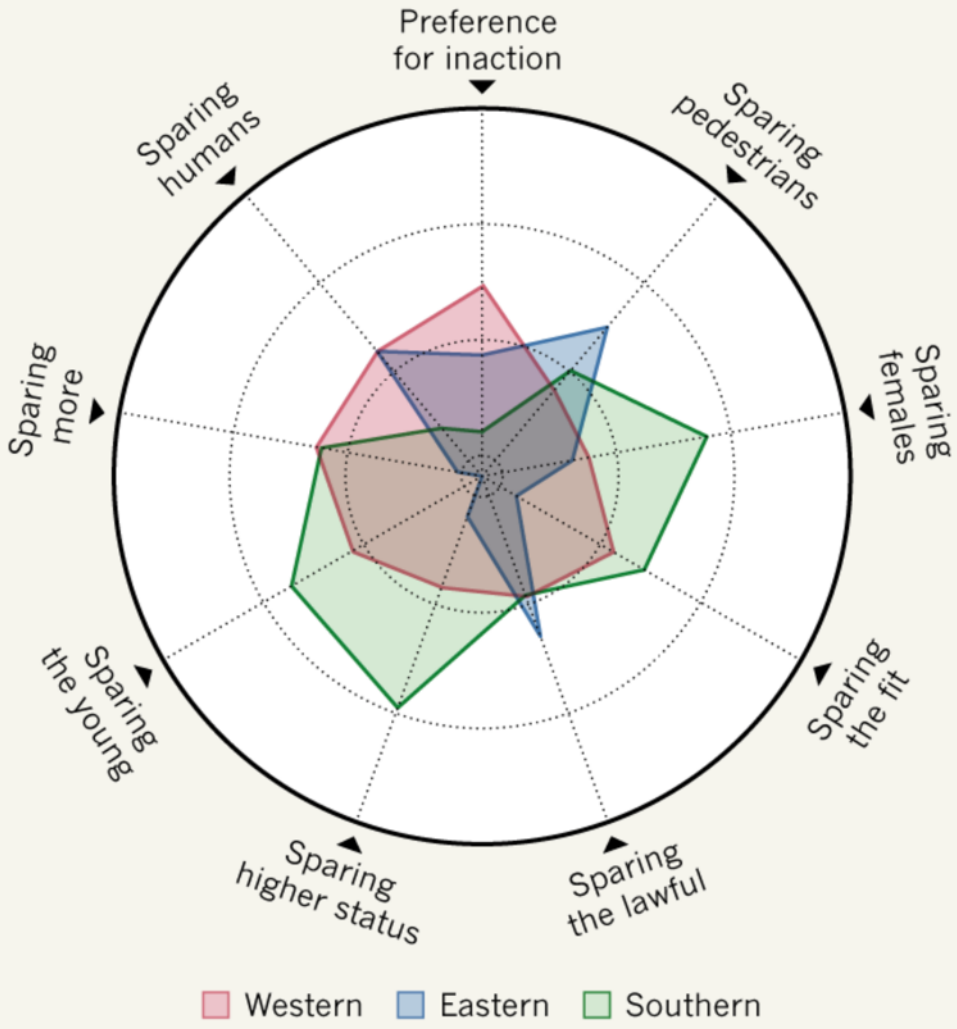
What should the self-driving car do?





moralmachine.org, 2,3 Mio. Teilnehmer, 13 Situationen

Es gibt keine universell gültigen moralischen Regeln.



Menschen die **sagen**, sie wollen Fußgänger schonen, würden ein solches Auto **nicht benutzen**.

Ethik: eine Frage der Verantwortung

Was sollen wir tun?

aktiv, proaktiv

Wer ist schuld?

passiv, retrospektiv

Entwickler? Experten?
Der Hersteller?
Das Auto?

Kriterien für Schuldfähigkeit

Fehlverhalten

Ein Akteur hat einen **Verstoß** gegen ein Gesetz oder allgemein akzeptierte Regeln begangen.

Kausalzusammenhang

Handlung oder Unterlassung eines Akteurs

Notwendige Voraussetzung für das Ergebnis

Vorhersehbarkeit

Konnte der Akteur die **Folgen** seiner Handlung **absehen**?

Akteure müssen sich *angemessen* informieren

Handlungsfreiheit

Akteur konnte aus **Alternativen** wählen

Akteur wurde **nicht gezwungen oder überlistet**.

Ibo van de Poel, Lambèr Royakkers (2011). Ethics, Technology, and Engineering: An Introduction. Wiley Blackwell.

A Simple Sticker Tricked Neural Networks Into Classifying Anything as a Toaster

Melanie Ehrenkranz
1/03/18 11:05am • Filed to: DEEP LEARNING

48.4K 31 2

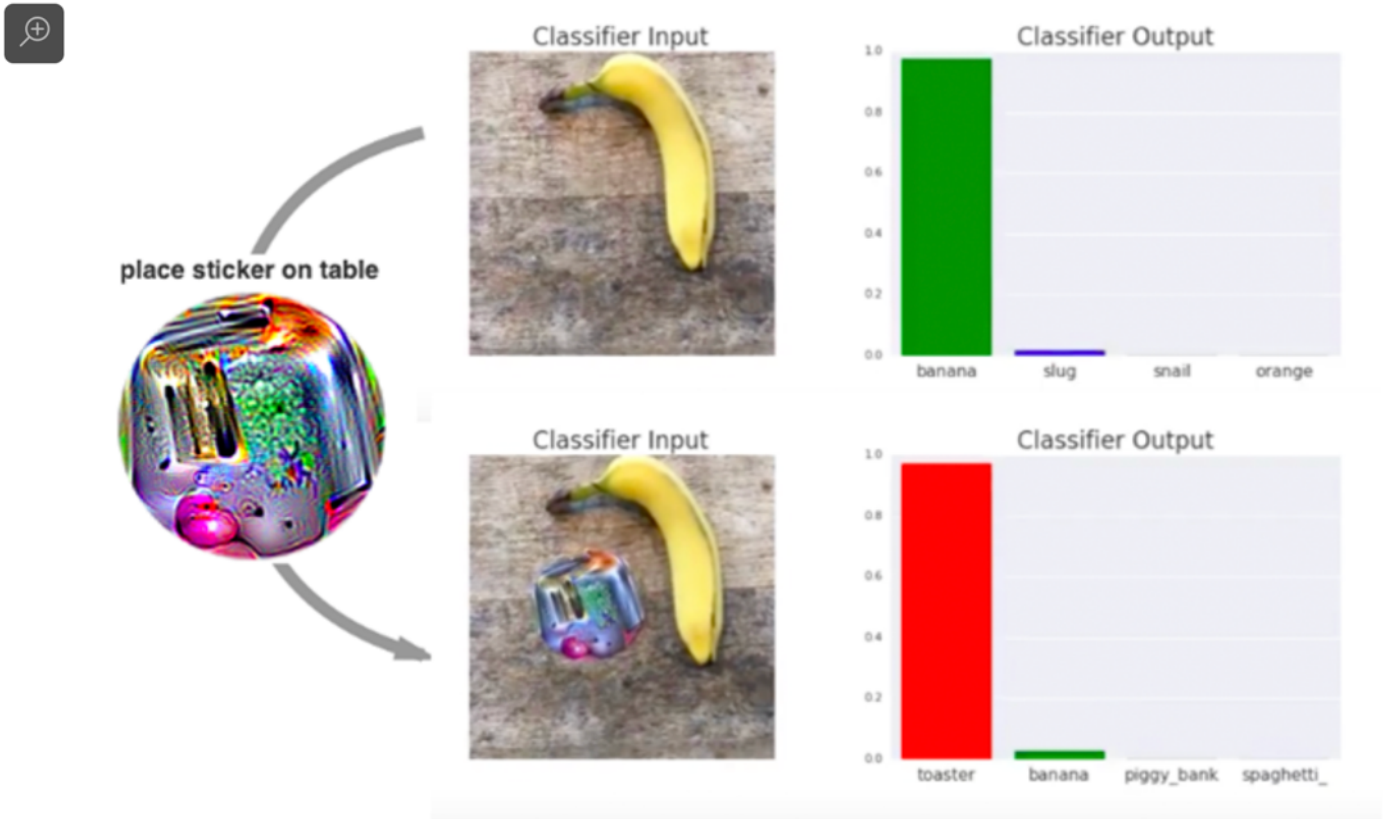
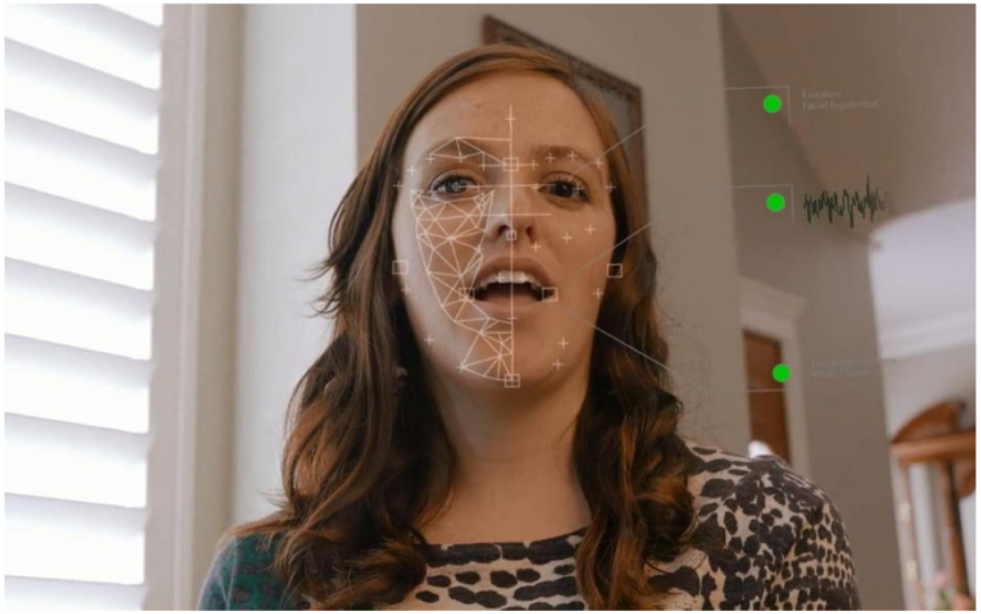


Image: Tom B. Brown/Dandelion Mané

AI erstmals bei Vorstellungsgesprächen in Großbritannien eingesetzt, um die besten Bewerber zu finden



An applicant being interviewed on their phone

By **Charles Hymas**

27 SEPTEMBER 2019 • 10:00 PM

350 Merkmale

Sprachgebrauch, verbale Fähigkeiten, passiv vs aktiv, ich vs wir, Wortwahl, Satzlänge, Tonfall, Geschwindikgeit, Empathie

Das soll funktionieren?

Der „Markt“ glaubt jedenfalls daran.

Vendor name	Funding	# of employees	Location
8 and Above	–	1-10	WA, USA
ActiView	$6.5M	11-50	Israel
Assessment Innovation	$1.3M	1-10	NY, USA
Good&Co	$10.3M	51-100	CA, USA
Harver	$14M	51-100	NY, USA
HireVue	$93M	251-500	UT, USA
impress.ai	$1.4M	11-50	Singapore
Knockri	–	11-50	Canada
Koru	$15.6M	11-50	WA, USA
LaunchPad Recruits	£2M	11-50	UK
myInterview	$1.4M	1-10	Australia
Plum.io	$1.9M	11-50	Canada
PredictiveHire	A$4.3M	11-50	Australia
pymetrics	$56.6M	51-100	NY, USA
Scoutible	$6.5M	1-10	CA, USA
Teamscope	€800K	1-10	Estonia
ThriveMap	£781K	1-10	UK
Yobs	$1M	11-50	CA, USA

https://arxiv.org/abs/1906.09208

Realität?

James Ball ✓
@jamesrbuk

Vision: algorithms will make hiring better as they don't discriminate

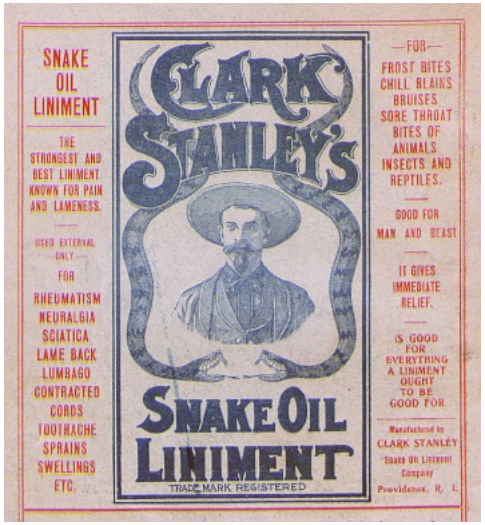
Reality: "One HR employee for a major technology company recommends slipping the words "Oxford" or "Cambridge" into a CV in invisible white text, to pass the automated screening."

How to persuade a robot that you should get the job
theguardian.com

13:16 · 04.03.18 · Twitter for iPhone

„Die meisten KI-Produkte sind Schlangenöl.“

Arvind Narayanan (Princeton)



Warum glauben so viele an die Wirkung?

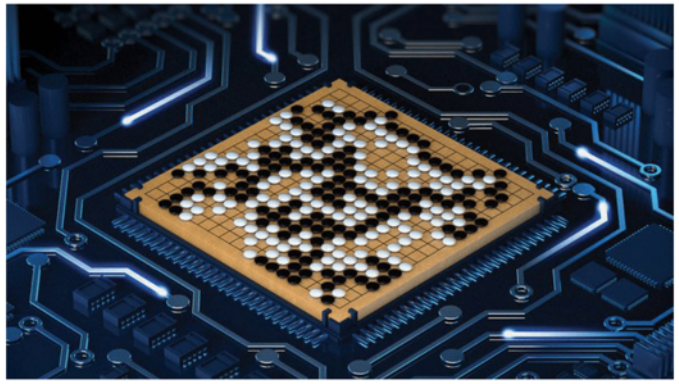
KI

griffiger Oberbegriff für viele Technologien

ML Data Science Statistik Automatisierung

Punktuell tatsächlich bemerkenswerte Fortschritte
(aber: für „künstliche“ Probleme)

Künstliche Intelligenz: AlphaGo Zero übertrumpft AlphaGo ohne menschliches Vorwissen



Produkte mit „KI“-Label verkaufen sich einfach besser.

Beeindruckende Fortschritte

Inhalte wiedererkennen (Shazam, Bildersuche)

Gesichtserkennung *

Krankheiten diagnostizieren

Audio-Transkription (Speech-to-Text)

Spracherkennung

Deepfakes *

* Ethische Bedenken wg. hoher Genauigkeit

Wahrnehmung

Nicht perfekt, aber es geht voran

Spam erkennen

Urheberrechtl. gesch. Material erkennen

Essays autom. bewerten

Hate Speech erkennen

Passende Inhalte vorschlagen

Ethische Bedenken da Fehler unvermeidlich sind.

Entscheidungen automatisieren

Fundamental schwierig

Rückfallquoten von Straftätern vorhersagen

Arbeitsleistung vorhersagen

Terror-Risiken vorhersagen

Predictive Policing

Gefährdete Kinder, Studierende, … erkennen

Ethische Probleme werden durch Ungenauigkeit verstärkt.

Sozialverhalten vorhersagen

Unvollständig und ungenau aber hilfreich zur Erläuterung

Wahrnehmung

beeindruckende Fortschritte

Entscheidungen automatisieren

nicht perfekt aber es geht voran

Sozialverhalten vorhersagen

fundamental schwierig

Schlangenöl

BEISPIEL #1

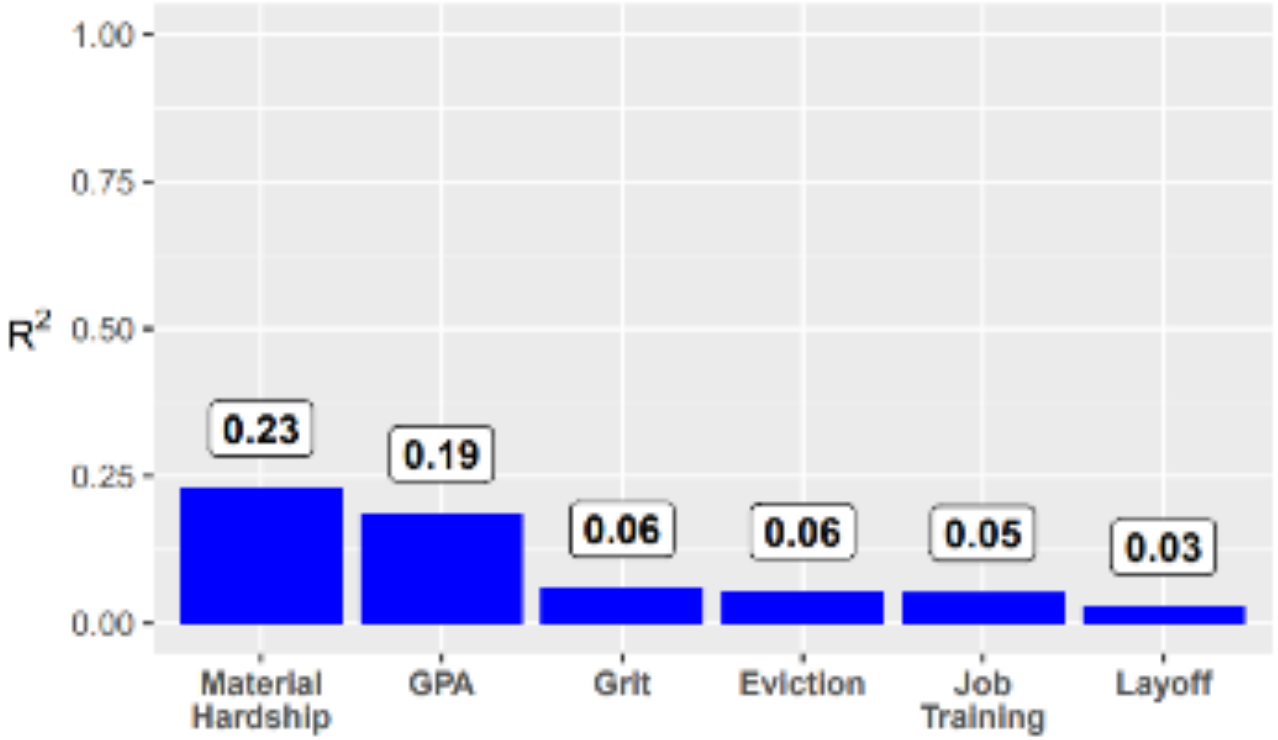
Fragile Families
& Child Wellbeing Study
PRINCETON | COLUMBIA

Lassen sich *Social Outcomes* mit KI vorhersagen?

457 Wissenschaftler
4242 Familien
12.942 Merkmale

https://fragilefamilies.princeton.edu/about

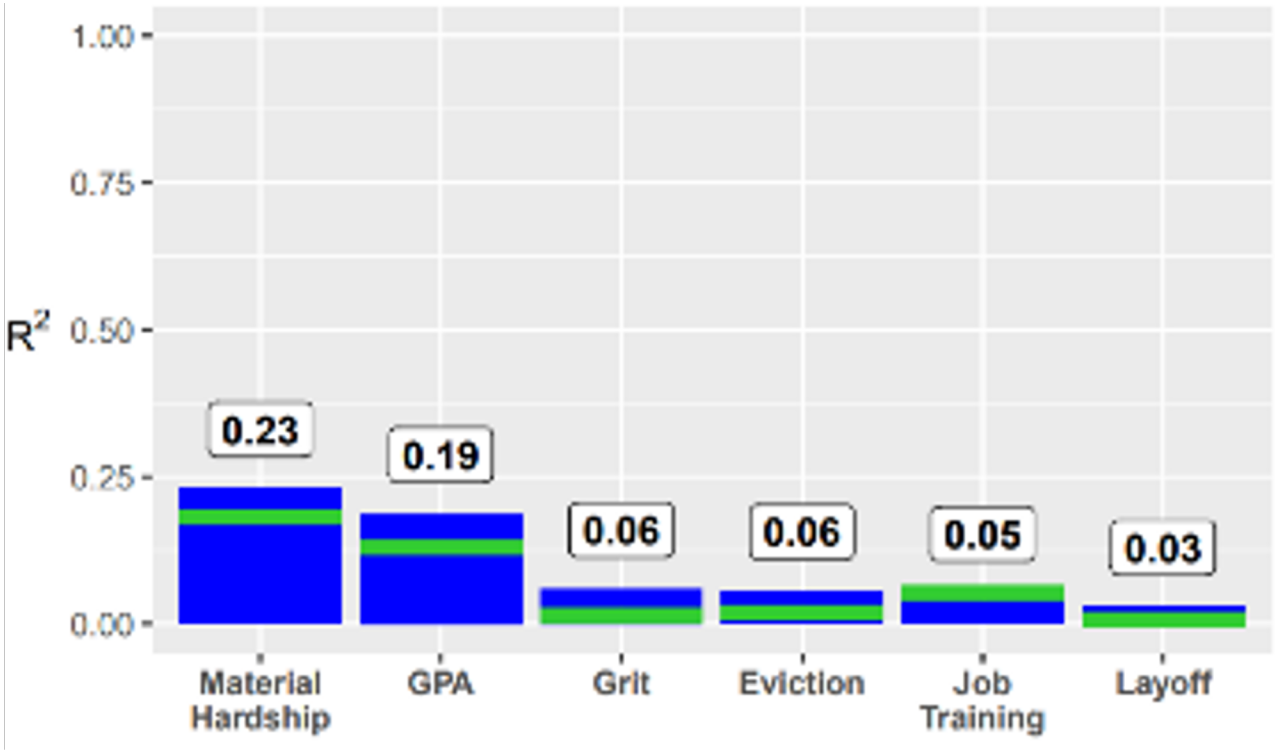
Genauigkeit der Vorhersage mit dem besten KI-Modell



R^2 = 1: exakte Vorhersage welche Eigenschaften mit Social Outcomes korrelieren

R^2 = 0: alle Eigenschaften korrelieren gleichermaßen mit Social Outcomes

Grüne Linie: Lineare Regression mit 4 Variablen



BEISPIEL #2

COMPAS

Lässt sich vorhersagen ob verurteilte Straftäter rückfällig werden?

COMPAS	137 Features	65% ± 1%
Logist. Regression	2 Features	67% ± 2%

Alter und
Vorstrafen

Dressel & Farid. *The accuracy, fairness, and limits of predicting recidivism*. Science Advances 2018.

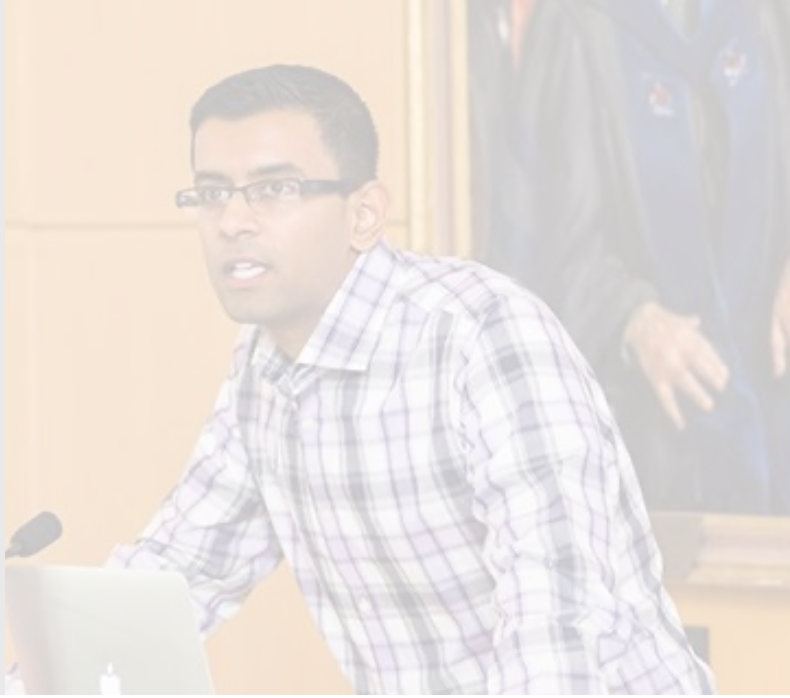
Vorhersage von Social Outcomes: KI nicht wesentlich besser als manuelles Scoring.

Arvind Narayanan (Princeton)

Aber „KI“ hat handfeste Nachteile:

Sammlung sensibler Daten, aber Nutzen zweifelhaft	Machttransfer v. Domänen-experten zu Firmen
Unzureichende Erklärbarkeit	Naives Vertrauen in Objektivität u. Genauigkeit

Vorhersage von Social Outcomes: KI nicht wesentlich besser als manuelles Scoring.



Arvind Narayanan (Princeton)

Aber „KI“ hat handfeste Nachteile:

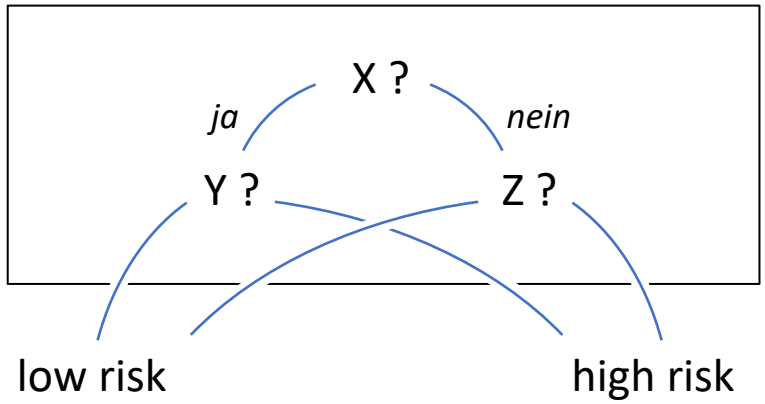
Sammlung sensibler Daten, aber Nutzen zweifelhaft	Machttransfer v. Domänen-experten zu Firmen
Unzureichende Erklärbarkeit	Naives Vertrauen in Objektivität u. Genauigkeit

Kann man mit ML frühzeitig das Sterblichkeitsrisiko bei einer Lungenentzündung erkennen?

Trainingsdaten:
Patientien mit Lungenentzündung
(gestorben und überlebt)

Black-Box-Modell
(Neuronales Netz)

Entscheidungsbaum



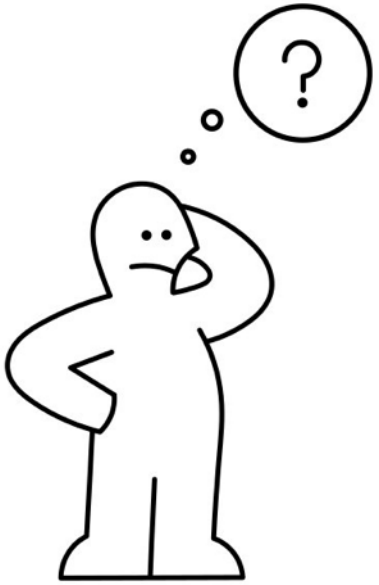
Caruana et al. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission, KDD 2015.

Kann man mit ML frühzeitig das Sterblichkeitsrisiko bei einer Lungenentzündung erkennen?

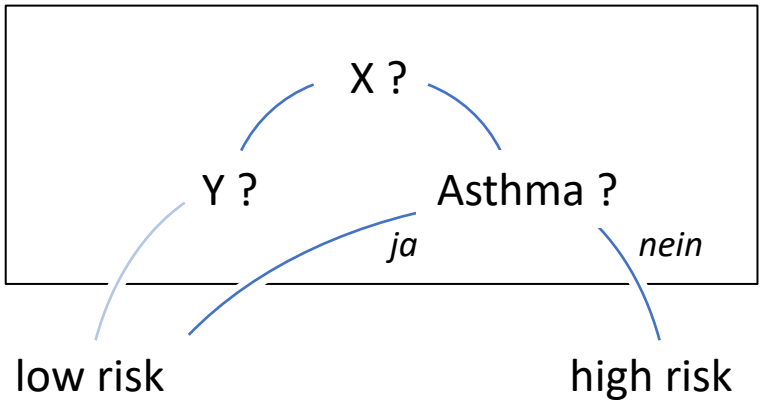
Trainingsdaten:
Patientien mit Lungenentzündung
(gestorben und überlebt)

Black-Box-
(Neuronale

low risk



Entscheidungsbaum



=> Explainable ML

Caruana et al. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission, KDD 2015.

Der Ruf nach

Algorithmentransparenz

Wenn ein Unternehmen „Algorithmen“ einsetzt,
sollte es verpflichtet werden ...

... den Mechanismus zu erklären?

Prosa, Code, Beispiele für Ein- und Ausgaben?

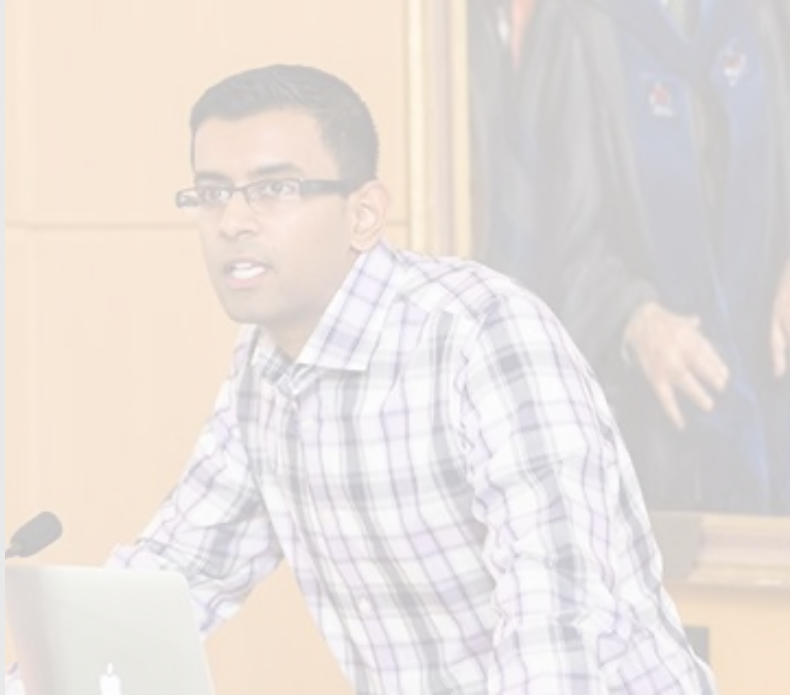
... die Trainingsdaten veröffentlichen?

Kann die Privatsphäre von Personen verletzen.

... mit Testergebnissen die Richtigkeit belegen?

Wer setzt die Standards, wie das zu tun ist?

Vorhersage von Social Outcomes: KI nicht wesentlich besser als manuelles Scoring.



Arvind Narayanan (Princeton)

Aber „KI“ hat handfeste Nachteile:

Sammlung sensibler Daten, aber Nutzen zweifelhaft

Machttransfer v. Domänen-experten zu Firmen

Unzureichende Erklärbarkeit

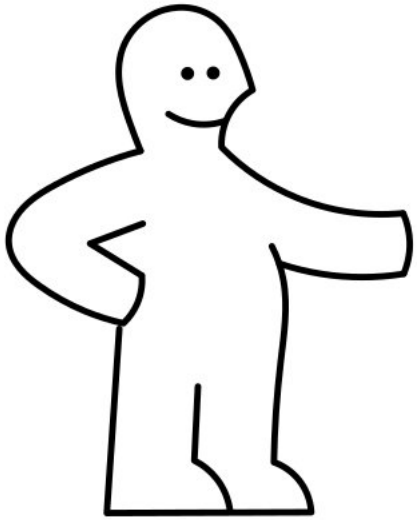
Naives Vertrauen in Objektivität u. Genauigkeit

GLOBAL TIMES

Professor claims AI can spot criminals by looking at photos 90 % of the time

By Liu Xin Source:Global Times Published: 2017/1/3 19:43:39

MIT Technology Review



They take ID photos of 1856 Chinese men between the ages of 18 and 55 with no facial hair. Half of these men were criminals. They then used 90 % of these images to train a convolutional neural network to recognize the difference and then tested the neural net on the remaining 10 percent of the images.

The results are unsettling. Xiaolin and Xi found that the neural network could correctly identify criminals and noncriminals with an **accuracy of 89.5 %**. "These highly consistent results are evidences for the validity of automated face-induced inference on criminality, despite the historical controversy surrounding the topic," they say.

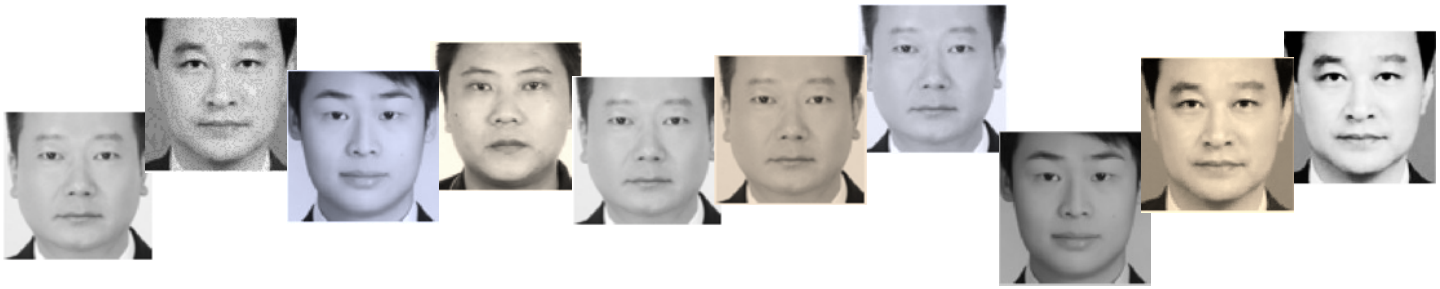
Comic People by @blinry

http://www.globaltimes.cn/content/1026928.shtml – https://www.technologyreview.com/s/602955/neural-network-learns-to-identify-criminals-by-their-faces/

90 % Accuracy

https://www.predictiveanalyticsworld.com/machinelearningtimes/accuracy-fallacy-the-medias-coverage-of-ai-is-bogus/10652/

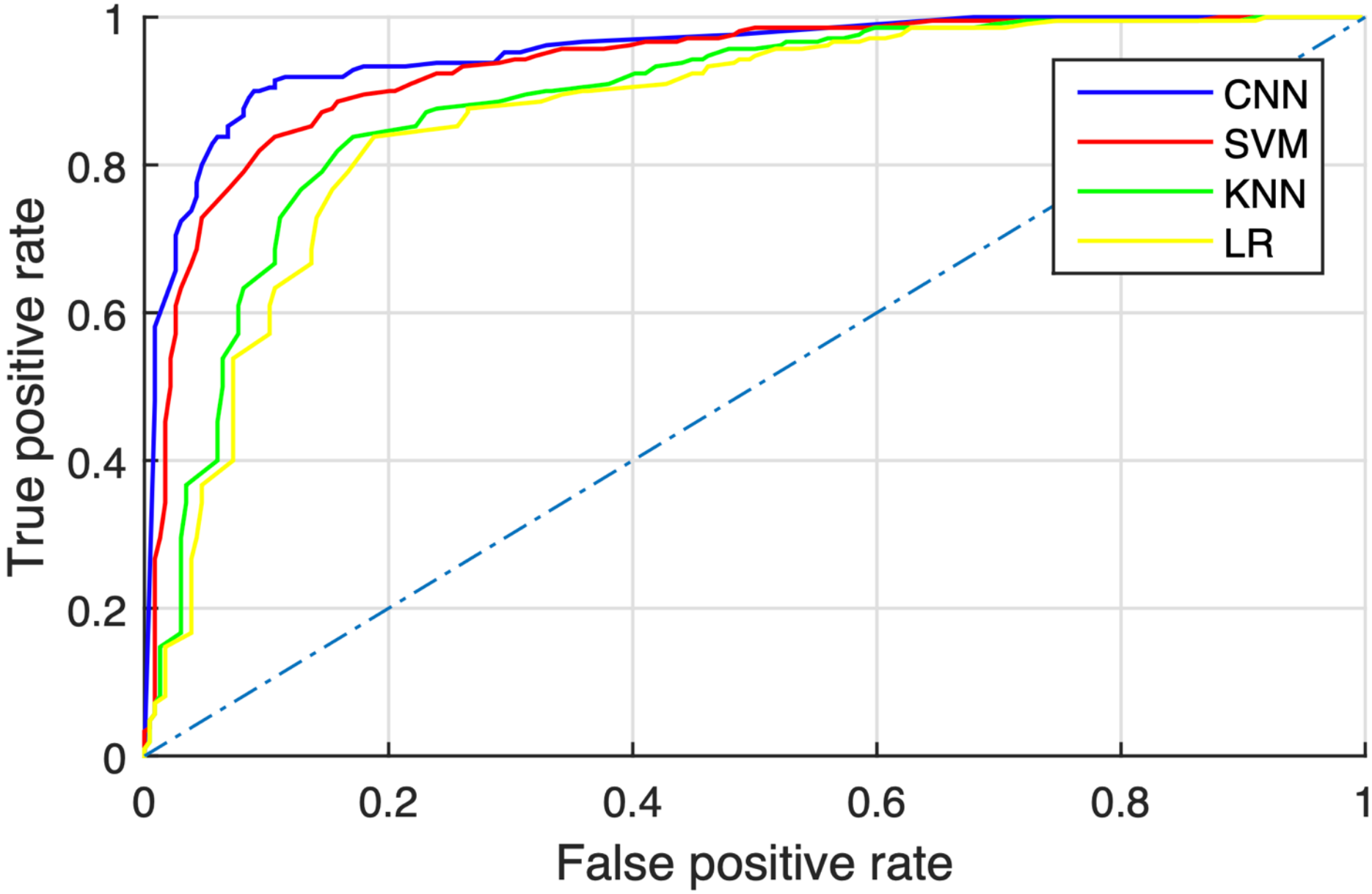
90 % Accuracy?



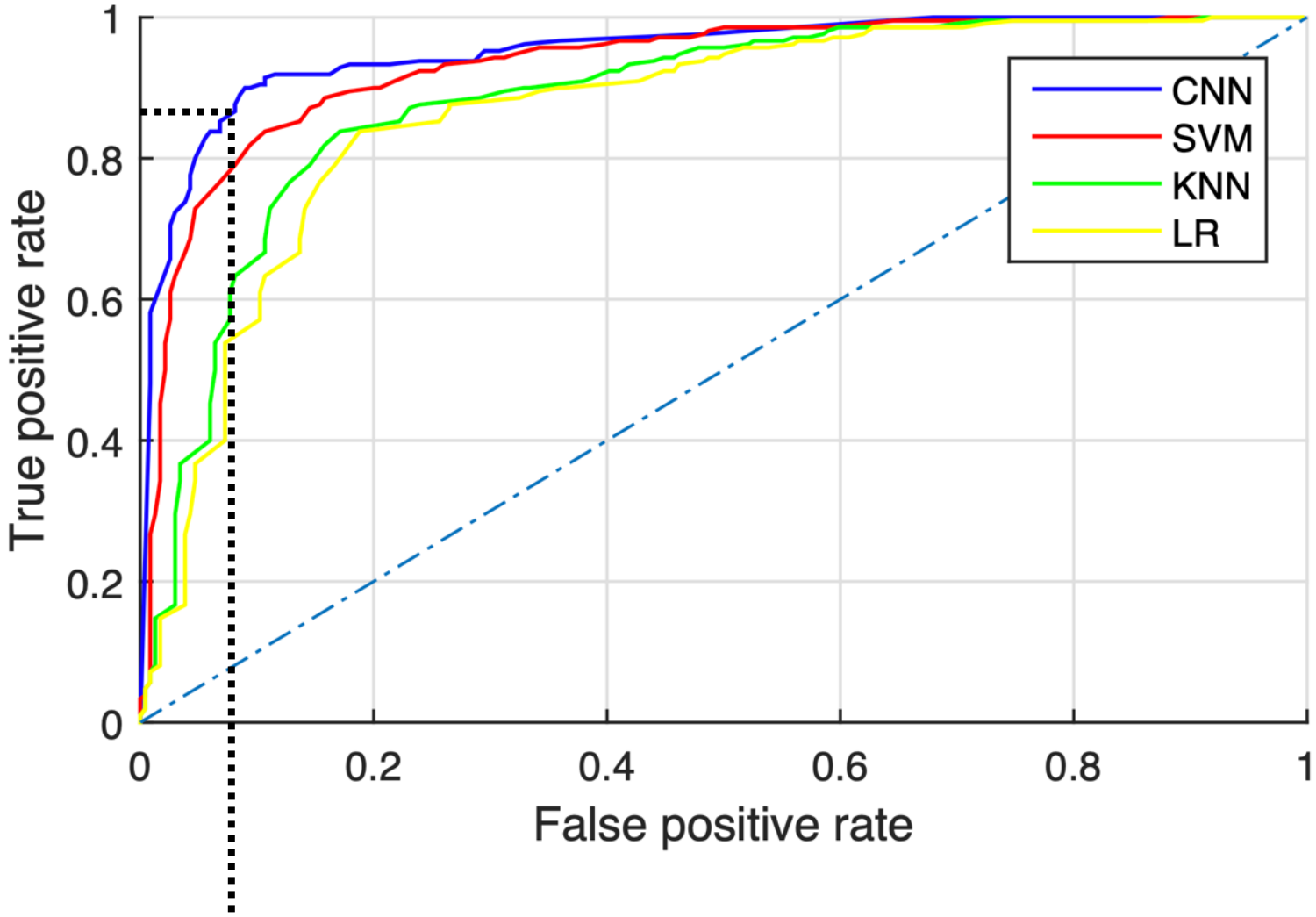
130 experiments, we examine the rate of correctly classifying a member of S into S_n or S_c, and then average the rates of each case over ten runs.

3.2. Results

In Figure 2, we plot the accuracies of all four classifiers in the above thirteen cases. As expected, the state-of-the-art CNN classifier performs the best, achieving 89.51% accuracy. The relatively high accuracy of CNN is also paralleled by all other three classifiers which are only few percentage points behind in the success rate of classification. These


True positive rate
False positive rate
0
0.2
0.4
0.6
0.8
1
CNN
SVM
KNN
LR

Kriminellen
übersehen: 17 %



Falscher Alarm: 7 %

Kriminellen übersehen: 17 % Falscher Alarm: 7 %

Straftäter in China: 164 je 100.000 Einwohner (0,16 %)

Unter 10.000 zufällig ausgewählten Bürgern also **16 Straftäter**.

Von 16 Straftätern werden ca. 3 übersehen, ca. 13 entlarvt.

Von den übrigen 9984 Nichtkriminellen
werden ca. 698 *als Straftäter* erkannt.

Wenn System ein Foto als „kriminell“ markiert …
nur **in 1,8% der Fälle** tatsächlich ein Krimineller.

= Base Rate Fallacy „Genauigkeit“ ist irreführend!

Drei für Entscheidung wichtige Features:

curvature of upper lip which is on average 23 % larger for criminals than for noncriminals;

the **distance between two inner corners of the eyes**, which is 6 % shorter;

and the angle between two lines drawn from the **tip of the nose** to the **corners of the mouth**, which is 20 % smaller.

https://www.technologyreview.com/s/602955/neural-network-learns-to-identify-criminals-by-their-faces/

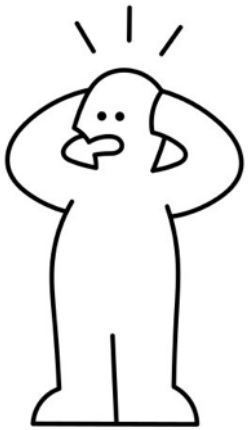
(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

Mugshots

gefunden mit Web-Crawler



KI und Datenschutz

Wie Betroffenenrechte durchsetzen?

Datenauskunft und Löschung

Machine Unlearning

Lucas Bourtoule*, Varun Chandrasekaran†, Christopher Choquette-Choo*, Hengrui Jia*,
Adelin Travers*, Baiwu Zhang*, David Lie*, Nicolas Papernot*§
University of Toronto*, Vector Institute§, University of Wisconsin-Madison†

***Abstract*—Once users have shared their data online, it is generally difficult for them to revoke access and ask for the data to be deleted. Machine learning (ML) exacerbates this problem because any model trained with said data may have memorized it, putting users at risk of a successful privacy attack exposing their information. Yet, having models unlearn is notoriously difficult. After a data point is removed from a training set, one often resorts to entirely retraining downstream models from scratch.**

We introduce SISA training, a framework that decreases the number of model parameters affected by an unlearning request and caches intermediate outputs of the training algorithm to limit the number of model updates that need to be computed to have these parameters unlearn. This framework reduces the computational overhead associated with unlearning, even in the worst-case setting where unlearning requests are made uniformly

personal data motivates us to examine how this right to be forgotten can be efficiently implemented for ML systems.

Because ML models potentially memorize training data [10], [11], it is important to sanitize models trained on data that have been deleted. This problem is tangential to privacy-preserving machine learning—enforcing differential privacy [12] does not solve it. While algorithms with differential privacy guarantee a bound on how much individual training points contribute to the model and ensure that this contribution remains small [13], [14], there remains *a non-zero* contribution from each point. If this was not the case, the model would not be able to learn at all from training data (see § III). In contrast, forgetting requires that *a*

Einsatz von KI schafft künstliche Probleme

Ethik und Datenschutz: bisher ungelöst

Social Outcomes vorhersagen: bisher Schlangenöl.

Genauigkeit: wenig aussagekräftig

Prof. Dr. **Dominik Herrmann**
Universität Bamberg

@herdom auf Twitter
https://dhgo.to/winfor-ki

Nick Craver
@Nick_Craver

"Uber will use artificial intelligence to identify drunk passengers. It will use factors like time of day, pickup location, and how long a user interacts with the app before making a decision."

That's not AI.
That's an if statement.

12:29 · 08.06.18 · TweetDeck

1.032 Retweets **2.820** Likes