

>>> Smart Borders?

>>> Wie die EU versucht Grenzübergänge mit einem diskriminierenden KI-Lügendetektor zu regulieren.

Name: verschiedene[†]

Date: September 12, 2021

[†]AG Link

>>> Inhalt

1. Was ist iBorderCtrl? (15min)

Akteure und Organisationsstruktur

Silent Talker

Geschichte des Lügendetektor

2. Grundlagen KI (15min)

Algorithmen, KI und Neuronale Netze

Overfitting

3. Grenzen von KI / Grenzen und Risiken von KI (20min)

KI und Bias

KI und Interpretierbarkeit

4. Bias Bingo mit / Bias in iBorderCtrl (20min)(interaktiv?)

5. Ausblick und Diskussion(20min)

Politische Einordnung

Wie und wofür forschen wir?

>>> Akteure und Organisationsstruktur

- * Horizon 2020 (auch Roborder)
- * Tresspass etc.
- * Finanzierung
- * Beteiligte Forschungseinrichtungen, beteiligte Unternehmen?
- * Aktueller Entwicklungsstand

>>> Silent Talker



Quelle: iborderctrl.eu

"The avatar is presented in a uniform to convey an air of authority." (K.Crockett et.al.)

>>> Die Geschichte des Lügendetektors - Der Polygraph

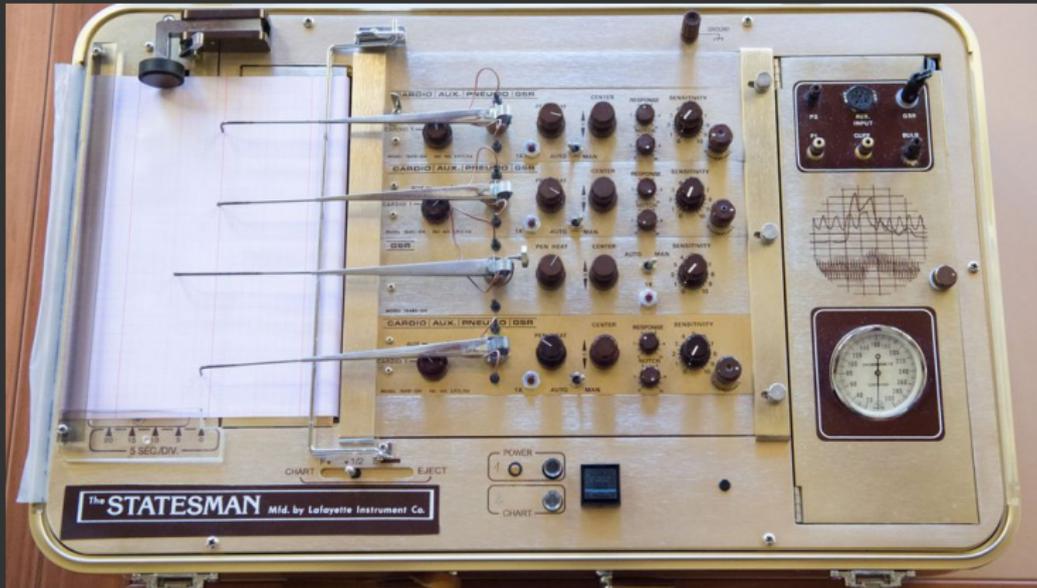


Foto: Sebastian Kahnert/ dpa

>>> Die Geschichte des Lügendetektors - Der Polygraph

- * Entwicklung in der ersten Hälfte des 20. Jahrhunderts
- * Maßgeblich von John A. Larson (Polizist und Physiologe) geprägt
- * Später patentiert und vermarktet
- * Forschung umstritten
- * Einige Studien legen nahe, dass Ergebnisse beinahe zufällig sind (Saxe, Ben-Shakhar, 1999)
- * Andere Systeme, die physiologische Reaktionen testen ähnlich umstritten

>>> Die Geschichte des Lügendetektors - Mikroexpressionen



Paul Ekman's Pictures of Facial Affect

>>> Die Geschichte des Lügendetektors - Mikroexpressionen

- * Facial Action Coding System (FACS) von Ekman und Friesen entwickelt,
- * 'Mikroexpressionen' codieren Gefühle des Menschen,
- * Annahme: Lügen ist emotionaler Akt und kann mit Mikroexpressionen entschlüsselt werden,
- * Studien zeigen: Vorgehen ist so exakt wie zufällige Vermutung,
- * Forschung ist uneinig, ob Mikroexpressionen bei jedem Menschen vorliegen und eindeutig sind.

>>> Die Geschichte des Lügendetektors - Die Rolle von KI

- * Versuche Mikroexpressionen und ähnliches mit KI zu klassifizieren,
- * Nutzung für die Optimierung von Werbeanzeigen (Facebook),
- * Projekt AVATAR (US-Projekt für Grenzkontrollen)

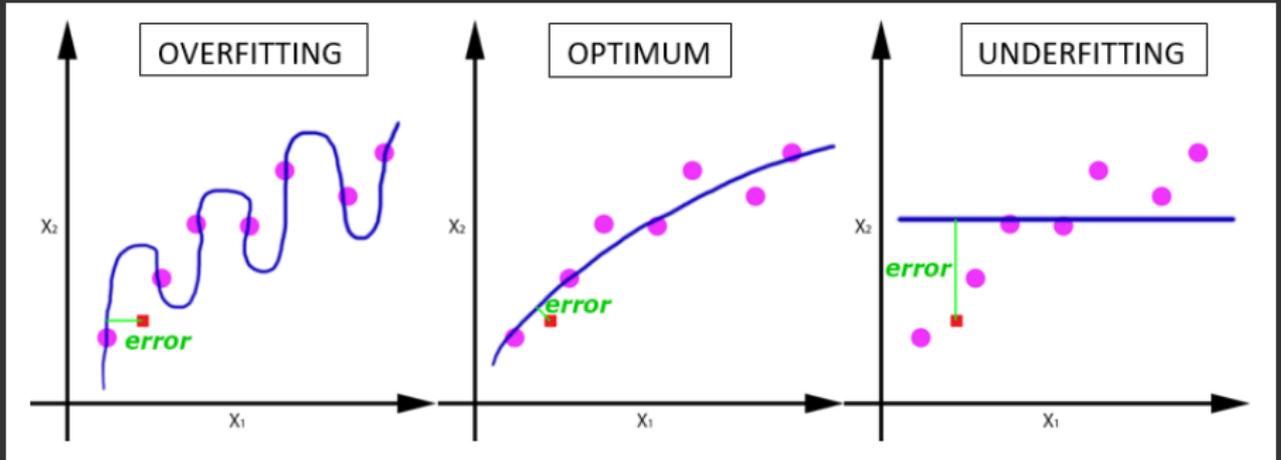
>>> Algorithmen, KI und Neuronale Netze

- * Algorithmus
- * KI / Machine Learning
- * Neuronale Netze und Deep Learning

>>> Aufbau der Datensätze

- * Supervised learning, Datenpaare
- * Training, Validation and Test Data

>>> Overfitting



Quelle: Sagar Sharma / Towards Data Science

>>> Was heißt interpretierbar?

- * Keine einheitliche Definition
- * Interpretierbarkeit als "Verstehen des Algorithmus":
 - * Bedeutung der Parameter kann verstanden werden
 - * Funktionsweise des Algorithmus kann für jeden Input nachvollzogen werden
 - * Kann bei KI oft nicht angewendet werden
- * Interpretierbarkeit durch methodische Analyse:
 - * Verschiedene Techniken
 - * Benutzung weiterer KI
 - * Analyse durch Manipulation von Datensätzen
 - * Konzepte sind neu
 - * Bedeutung und Korrektheit nicht geklärt

>>> Wofür braucht mensch Interpretierbarkeit?

- * Kausalität vs. Korrelation, was lernt die KI?
- * Bsp.: KI soll Basketbälle erkennen und wird mit Bildern unterschiedlicher Bälle trainiert. Erkennt die KI tatsächlich den Basketball oder einfach nur die Farbe Orange?
- * Wenn KI in gesellschaftlich relevanten Bereichen eingesetzt wird, kann es entscheidend sein Ergebnisse interpretieren zu können

>>> Missklassifikationshack

- * Ergebnisse der KI hängen stark vom Datensatz ab
- * Daraus ergibt sich ein Angriff auf KI's
- * Die Kriterien für die Entscheidung der KI stimmen nicht mit menschlichen Kriterien überein

>>> Panda oder Gibbon?

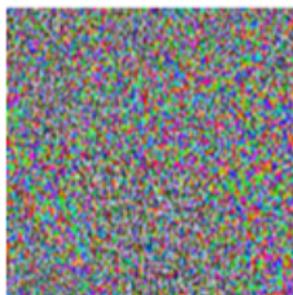


x

“panda”

57.7% confidence

+ .007 ×



$\text{sign}(\nabla_x J(\theta, x, y))$

“nematode”

8.2% confidence

=



$x +$

$\epsilon \text{sign}(\nabla_x J(\theta, x, y))$

“gibbon”

99.3 % confidence

Quelle: J. Goodfellow et. al.



M. BINY

Quelle:

>>> Arten von Bias

Bias = Verzerrung

- * Bias in den Daten
- * Bias durch Design des Algorithmus
- * Bias durch Rückkopplung im Gebrauch

Konsequenz: Diskriminierende Algorithmen (auch Gender Bias, Racial Bias, Neurodiversity Bias etc. genannt)

>>> Bias in den Daten

- * **Measurement Bias:** Wie werden bestimmte Eigenschaften gemessen/bestimmt?
 - * COMPAS: Verhaftungen auch von Familie etc. wurden genutzt um Risiko zu bewerten

>>> Bias in den Daten

- * Measurement Bias: Wie werden bestimmte Eigenschaften gemessen/bestimmt?
 - * COMPAS: Verhaftungen auch von Familie etc. wurden genutzt um Risiko zu bewerten
- * Omitted Variable Bias: Wichtige Daten werden nicht im Modell berücksichtigt
 - * Ein Programm das die Kündigung von Abos (zB Netflix) prognostiziert, aber neue Konkurrenzangebote nicht beachtet

>>> Bias in den Daten

- * Measurement Bias: Wie werden bestimmte Eigenschaften gemessen/bestimmt?
 - * COMPAS: Verhaftungen auch von Familie etc. wurden genutzt um Risiko zu bewerten
- * Omitted Variable Bias: Wichtige Daten werden nicht im Modell berücksichtigt
 - * Ein Programm das die Kündigung von Abos (zB Netflix) prognostiziert, aber neue Konkurrenzangebote nicht beachtet
- * **Representation Bias**: Fehlende Diversität in den verfügbaren Daten
 - * Coded Gaze, Joy Buolamwini

>>> Bias in den Daten

- * Measurement Bias: Wie werden bestimmte Eigenschaften gemessen/bestimmt?
 - * COMPAS: Verhaftungen auch von Familie etc. wurden genutzt um Risiko zu bewerten
- * Omitted Variable Bias: Wichtige Daten werden nicht im Modell berücksichtigt
 - * Ein Programm das die Kündigung von Abos (zB Netflix) prognostiziert, aber neue Konkurrenzangebote nicht beachtet
- * Representation Bias: Fehlende Diversität in den verfügbaren Daten
 - * Coded Gaze, Joy Buolamwini
- * **Aggregation Bias**: Spezifische Eigenschaften von Minderheiten gehen im gesamten Datensatz unter/Ableiten von Aussagen über Individuen aus Minderheit aus allgemeinem Datensatz
 - * Diskriminierungsklage gegen die Universität Berkeley

>>> Bias durch Design

- * Evaluation Bias: Ergebnis wird an unrealistischen Kriterien gemessen
 - * Bias in Gesichtserkennung fällt nicht auf, weil Benchmarkdatenset gebiast ist

>>> Bias durch Rückkopplung

* Beispiel Profiling

>>> Bias in iBorderCtrl

- * **Measurement Bias:** Schauspieler*innen statt echte Situationen, Stress beim Grenzübergang löst u.U. gleiche Symptome wie Lügen aus

>>> Bias in iBorderCtrl

- * Measurement Bias: Schauspieler*innen statt echte Situationen, Stress beim Grenzübergang löst u.U. gleiche Symptome wie Lügen aus
- * **Omitted Variable Bias:** Lügen/nicht Lügen zeigt sich in nichterfassten Phänomenen

>>> Bias in iBorderCtrl

- * Measurement Bias: Schauspieler*innen statt echte Situationen, Stress beim Grenzübergang löst u.U. gleiche Symptome wie Lügen aus
- * Omitted Variable Bias: Lügen/nicht Lügen zeigt sich in nichterfassten Phänomenen
- * **Representation Bias**: fehlende Diversität race, gender, neurdiversity, disability, health, scars

>>> Bias in iBorderCtrl

- * Measurement Bias: Schauspieler*innen statt echte Situationen, Stress beim Grenzübergang löst u.U. gleiche Symptome wie Lügen aus
- * Omitted Variable Bias: Lügen/nicht Lügen zeigt sich in nichterfassten Phänomenen
- * Representation Bias: fehlende Diversität race, gender, neurdiversity, disability, health, scars
- * **Aggregation Bias**: Reaktionen (z.B. Mikroexpressionen) nicht konsistent. Gleiches Verhalten kann unterschiedliches bedeuten.

>>> Bias in iBorderCtrl

- * Measurement Bias: Schauspieler*innen statt echte Situationen, Stress beim Grenzübergang löst u.U. gleiche Symptome wie Lügen aus
- * Omitted Variable Bias: Lügen/nicht Lügen zeigt sich in nichterfassten Phänomenen
- * Representation Bias: fehlende Diversität race, gender, neurdiversity, disability, health, scars
- * Aggregation Bias: Reaktionen (z.B. Mikroexpressionen) nicht konsistent. Gleiches Verhalten kann unterschiedliches bedeuten.
- * **Overfitting**: 73 Prozent in Testdaten vs 93 Prozent in Trainingsdaten, das impliziert Overfitting ist wahrscheinlich

>>> Bias in iBorderCtrl

- * **Evaluation Bias:** Testbedingungen entsprechen nicht den Einsatzbedingungen, zB Licht, Diversität

>>> Bias in iBorderCtrl

- * Evaluation Bias: Testbedingungen entsprechen nicht den Einsatzbedingungen, zB Licht, Diversität
- * **Bias durch Rückkopplung:** Wenn Einsatzdaten wieder eingespeist werden, dann: LD klassifiziert mehr Benachteiligte als lügend, Überprüfungsbeamt*in ebenso und gibt Daten zurück

>>> Testergebnisse

- * Anzahl verschiedener Personen in der Testdatenmenge: 1
- * Stichprobenstreuung: 24.37 (Truthful) / 34.29 (Deceptive)

Table IV: Classification Outcomes using Unseen Participants

Test No	Participant				Accuracy (%)	
	Truthful		Deceptive		Truthful	Deceptive
	Gender	Ethnicity	Gender	Ethnicity		
1	M	EU	M	A/A	100	57
2	M	A/A	F	EU	50	36
3	M	A/A	F	EU	50	100
4	M	EU	F	EU	90	100
5	M	A/A	M	EU	100	10
6	M	EU	M	EU	72	100
7	M	A/A	F	EU	100	100
8	F	EU	F	A/A	38	100
9	M	EU	M	EU	80	60
Overall Accuracy (%)					75.55	73.66

>>> Politische Einordnung

- * Unfreiwillige Datenerhebung zur "Verbesserung" des Algorithmus
- * Entwicklung Ethischer Normen für KI (EU KI Standards, Gesellschaft für Informatik)

>>> Wie und wofür forschen wir?

- * Instrumentalisierung von Wissenschaft
- * Was ist mein wissenschaftlicher Standard und worauf gründet er?
- * Gibt es Ziele und Werte die nicht von der Wissenschaft vorgegeben werden, sondern die wir uns selbst setzen müssen?