Künstliche Intelligenz → verständlich
Technische Hochschule Wildau

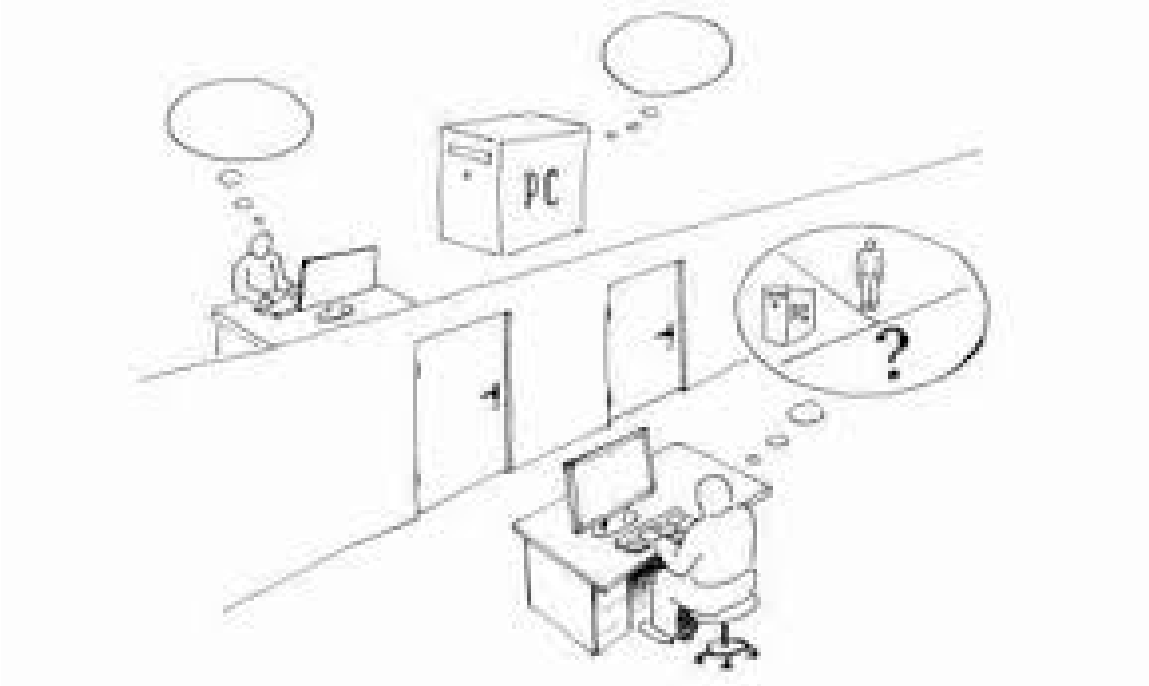
Können Maschinen ethisch handeln?

Prof. Dr. David Scheffer
CAPTA Institut
„Computer Aided Psychometric Text Analysis“

Der Turing-Test: Generelle, Starke KI versus Spezifische, Schwache KI

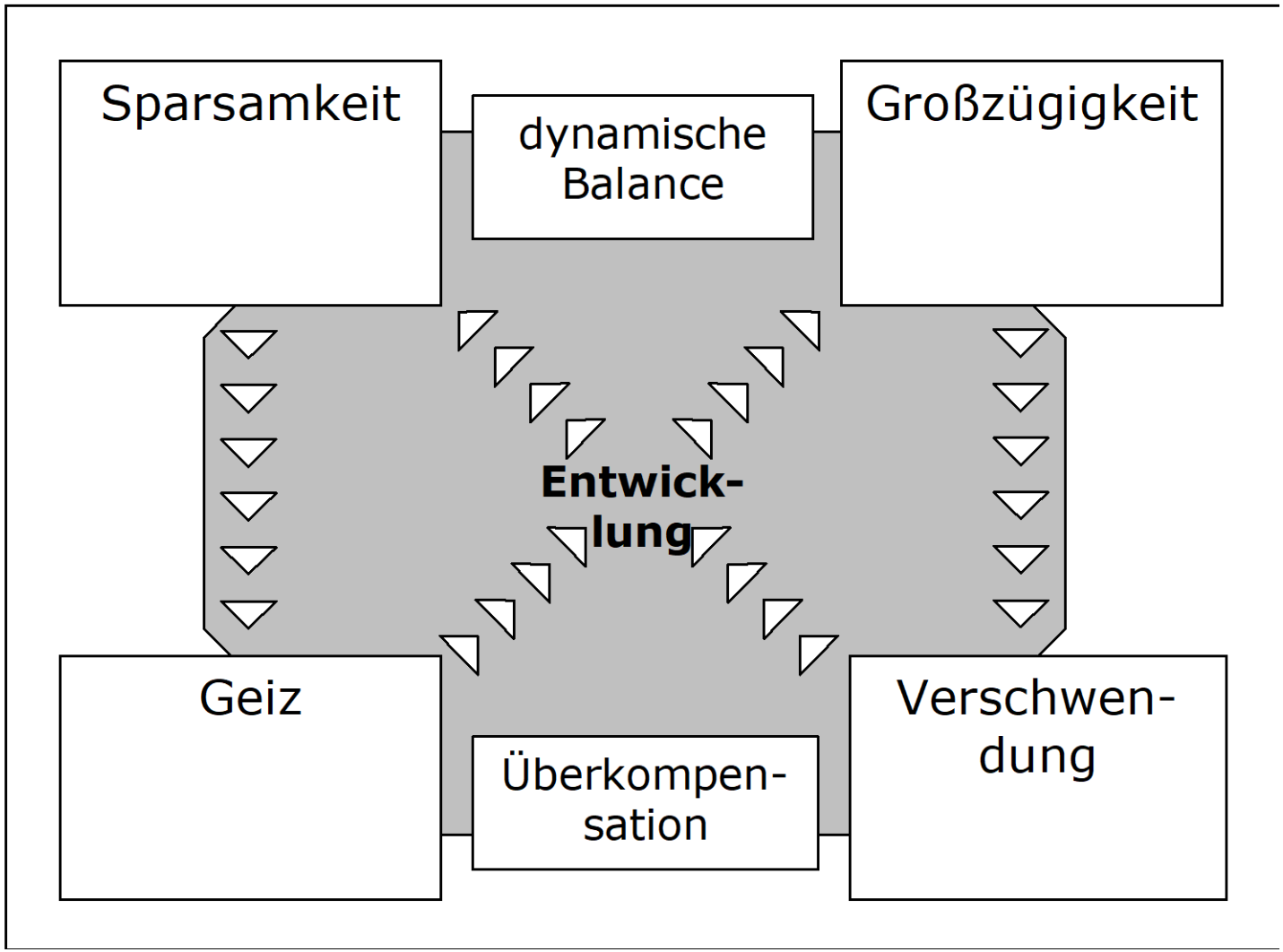
NORDAKADEMIE
HOCHSCHULE DER WIRTSCHAFT

Maschinen können so konstruiert werden, dass sie das Verhalten des menschlichen Geistes in eingegrenzten Bereichen sehr realistisch simulieren (schwache KI).



- Turing, A.M. (1950). Computing Machinery and Intelligence. *Mind. A Quarterly Review of Psychology and Philosophy, 236.*

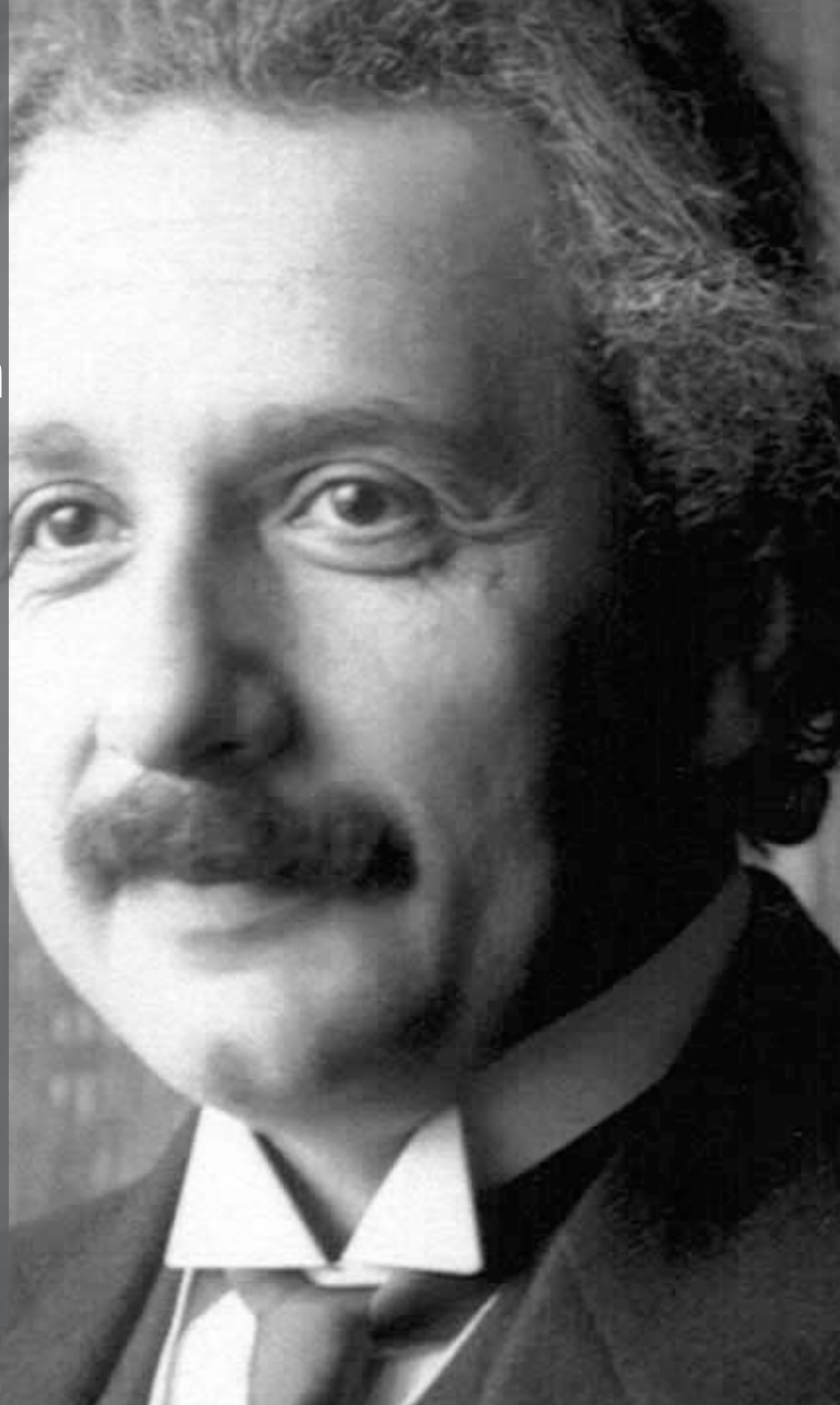
Ethik als Wertequadrat (Aristoteles, Schulz v. Thun)



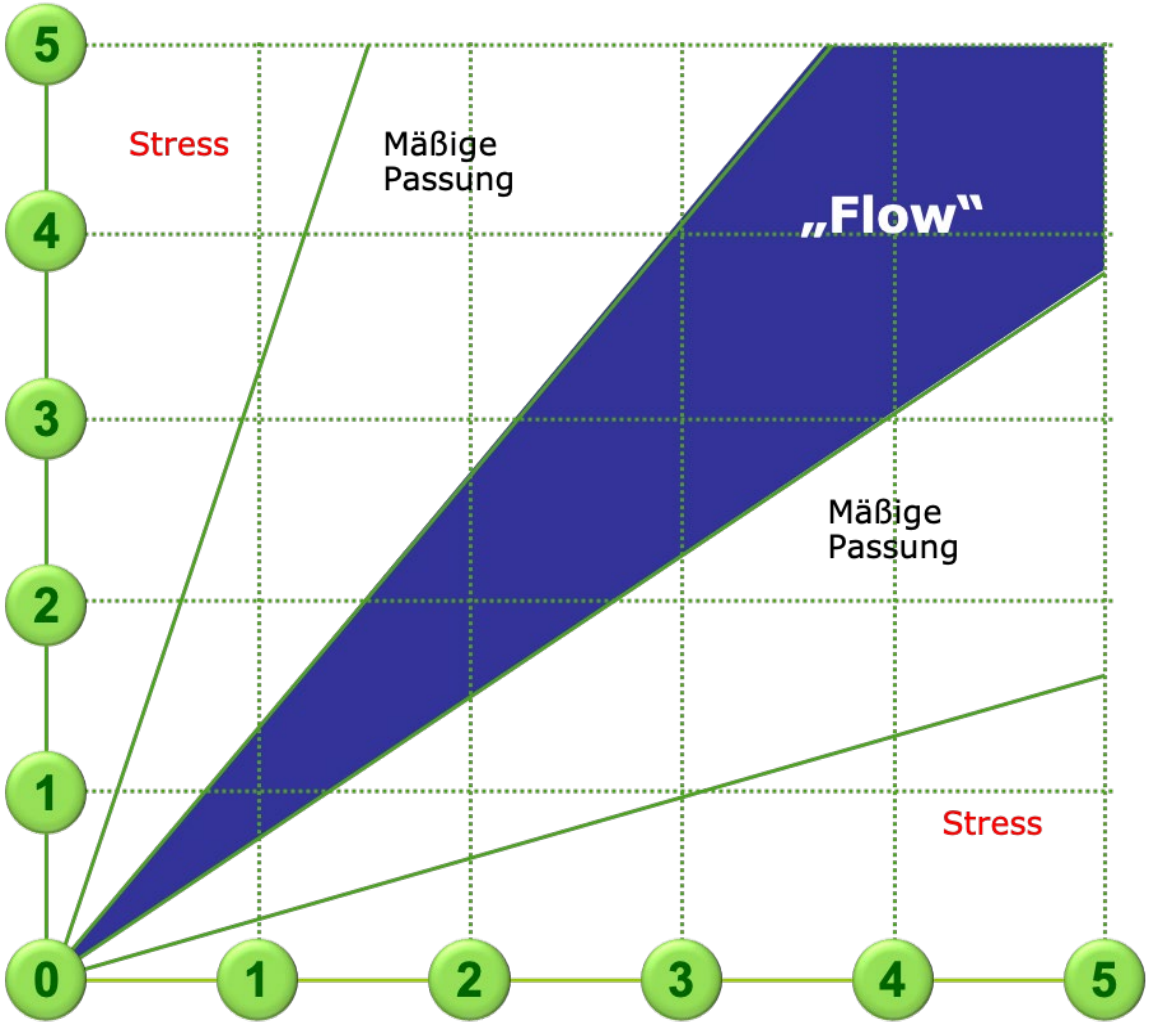
„Abschlussnoten sind wertlos bei der Personalauswahl. Wir haben festgestellt, dass sie rein gar nichts vorhersagen.“

Das Zitat stammt nicht von frustrierten Bewerbern mit schlechten Uni-Zeugnissen, sondern von Laszlo Bock, dem Personalchef von Google. (aus: Human Resources Manager, Ausgabe Februar/März 2016).

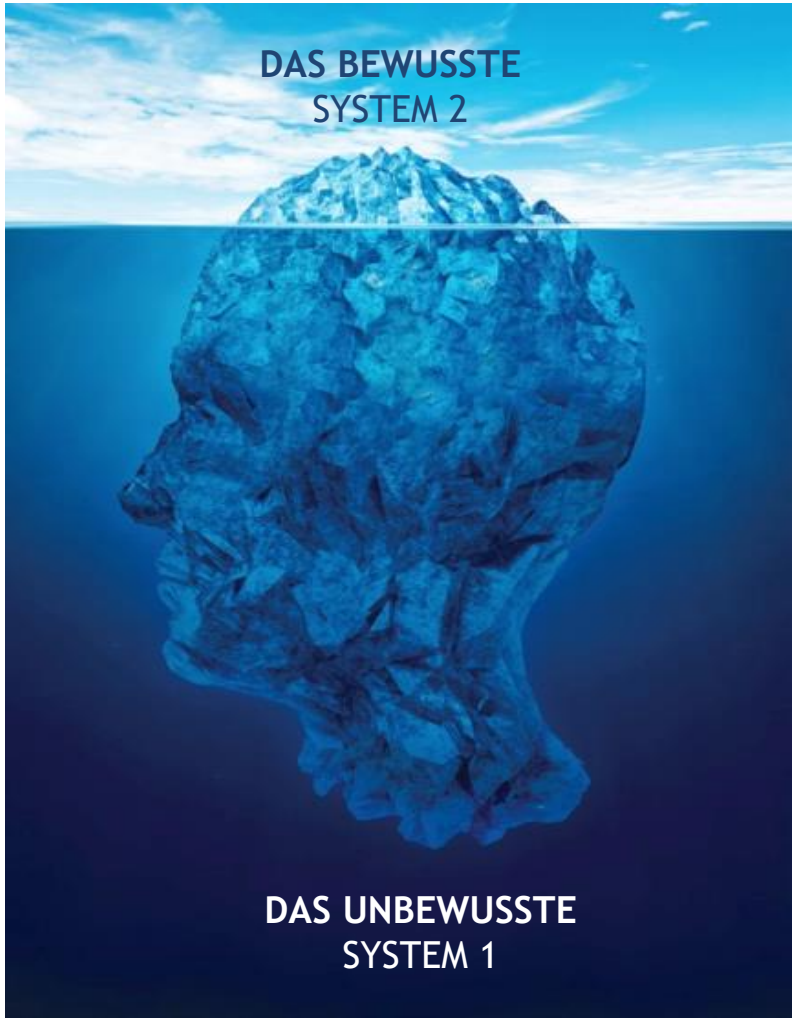
„Der Zusammenhang von Noten und Ausbildungserfolg ist nach einer Metaanalyse mit r = 0,41 gut; er sinkt aber für allgemeinen Berufserfolg auf Werte, die durchweg unter r = 0,20 liegen“ (Schwarzinger, Frintrup & Spengler, personalmagazin 09 / 14).



Person-Job-Team-Organisations-Passung erzeugt „Flow“

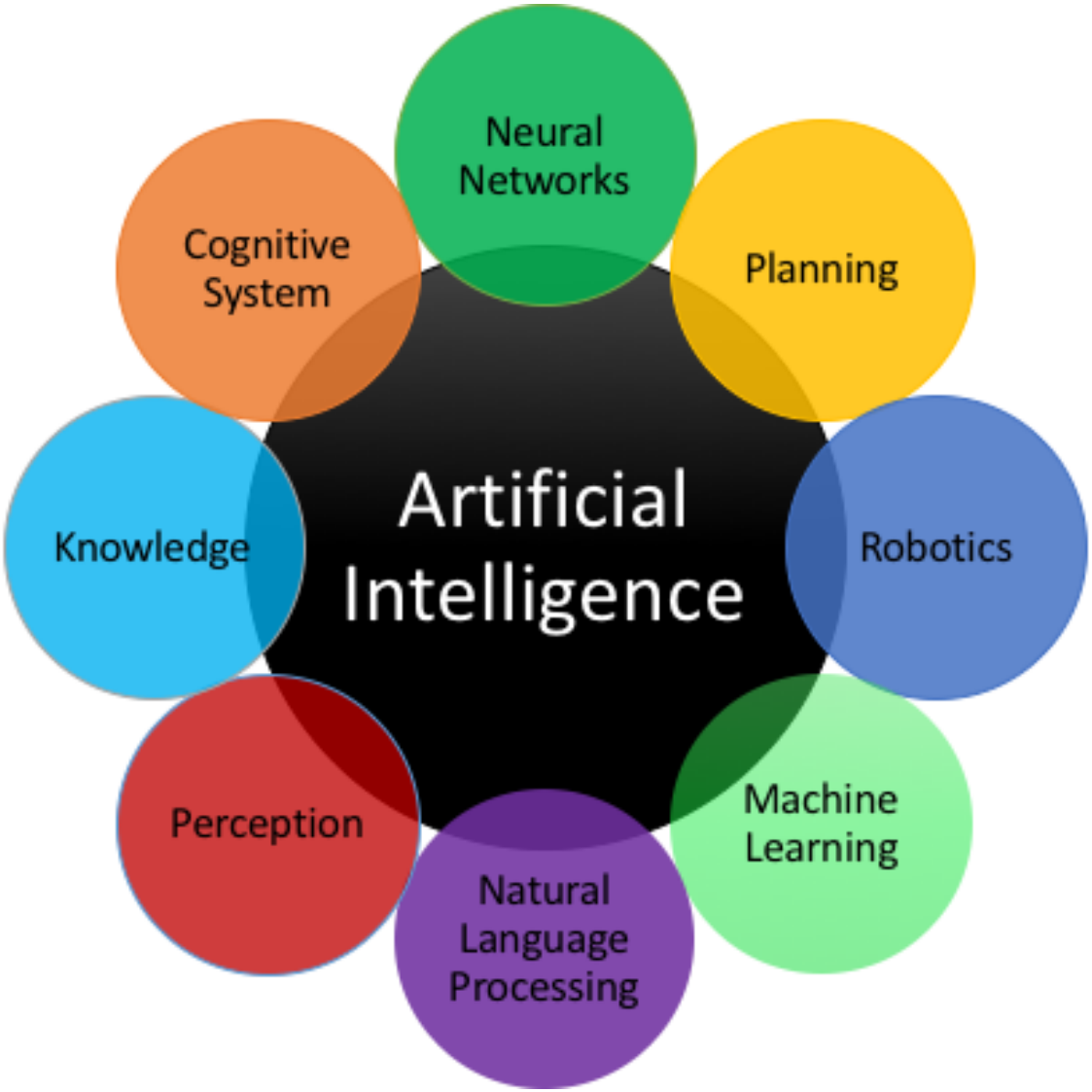


Existenz des Unbewussten



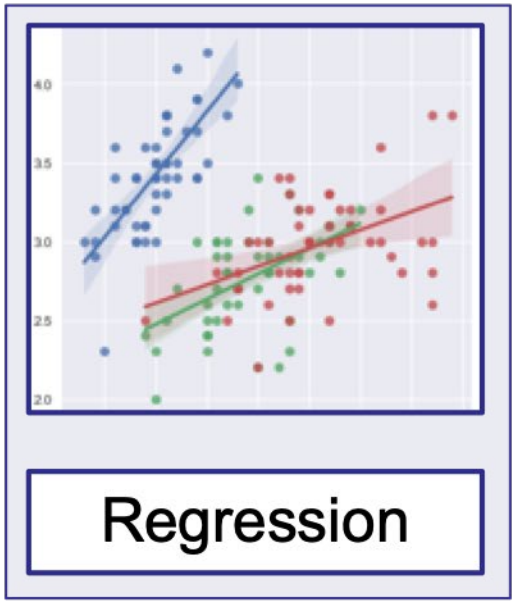
- Schnelles Denken und langsames Denken (Daniel Kahneman)
- Die Persönlichkeit ist die Brille, durch die wir wahrnehmen, entscheiden, handeln. Sie ist aber zu weiten Teilen unbewusst
- Unethisches Verhalten kann aus unbewussten internen und externen Konflikten entstehen

Künstliche Intelligenz (KI)

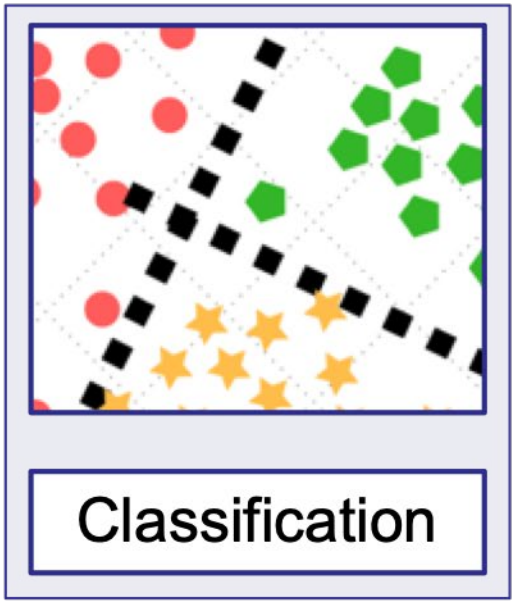


Quellen: Scoro.com; RazorRobotics.com

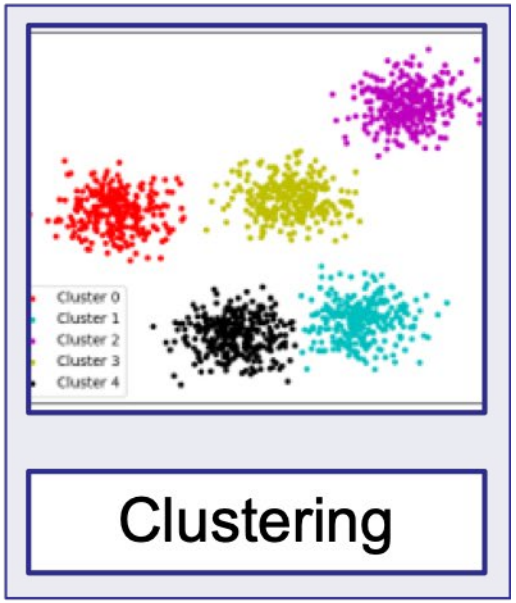
Aufgaben des Maschinen-Lernens



Abhängig Variable durch unabhängige Variable(n) erklären



Festgelegte, bereits bekannte Klassen (Instanz-Zuordnung)



Gruppierung unbekannter Objekte zur Musterfindung

Quellen: pydata.org, nlpforhackers.io, imperva.com.

Amazon trainiert KI an Bewerberdaten – es waren fast nur Männer

NORDAKADEMIE
GRADUATE SCHOOL

In Bündnissen wie „Partnership on AI“ oder „Open AI“ entwerfen die kalifornischen Techfirmen ethische Regeln für künstliche Intelligenz. Sie setzen sich für fehler- und verzerrungsfreie Software ein.

Das AGG (Allgemeines Gleichbehandlungsgesetz) setzt weitere Grenzen der Willkür im Umgang mit Person bezogenen Daten, dazu zählt das Verbot jeglicher Benachteiligung von Mitarbeitern und Bewerbern aus Gründen der Rasse, der ethnischen Herkunft, des Geschlechts, der Religion oder Weltanschauung, einer Behinderung, des Alters oder der sexuellen Identität.

Lösungsansatz

- Syntax: „Frauen sind“ impliziert, dass es eine wahre zutreffende Antwort für alle Frauen gibt, ausnahmslos. Und das ist falsch. Das gleiche gilt für ethnische Gruppen. Aufgrund des AGG darf Geschlecht etc. kein Kriterium bei der Personalauswahl sein.
- Datenquellen müssen repräsentativ sein, die Sichprobentheorie muss beachtet werden. Aber auch weitere Einflussparameter wie Abteilung, Vorgesetzte etc. müssen vorher geklärt werden. Auch: Psychopathie, dunkle Triade.
- Herkunft und Qualität vorhandener Bewertungen kritisch prüfen

Entgegen jeder Erwartung

Etablierte eignungsdiagnostische Verfahren gibt es bei Talanx seit Jahrzehnten. Den Ergebnissen wird vertraut, dennoch hat der Konzern ein neues Instrument eingeführt. Der Antrieb war: reine Neugierde. Im Raum stand die Behauptung, dass ein kurzer Dialog mit einem Computer zu einer validen Messung von Persönlichkeitsmerkmalen und Fähigkeiten einer Führungskraft führen könne. Niemand glaubte ernsthaft an einen Erfolg des Versuchs. Auf der anderen Seite ging es nur um einen Aufwand von wenigen Minuten. Minuten, die das Ende der Assessment-Center bei Talanx einläuten sollten. Als Testpersonen stellten sich die Vorstände zur Verfügung. Entgegen jeder Erwartung überzeugten die Resultate. Heute werden konzernweit alle leitenden Angestellten mithilfe dieses Verfahrens eingestellt und befördert.

Schneller

Precire Technologies ermittelt anhand einer 10- bis 15-minütigen Sprachprobe ein Führungsprofil des Teilnehmers – schneller, als es die Vorstellungskraft für möglich hält. Der Zeitaufwand gegenüber etablierten Einzel- oder Gruppen-Assessment-Centern verringert sich um mehr als 90 Prozent.

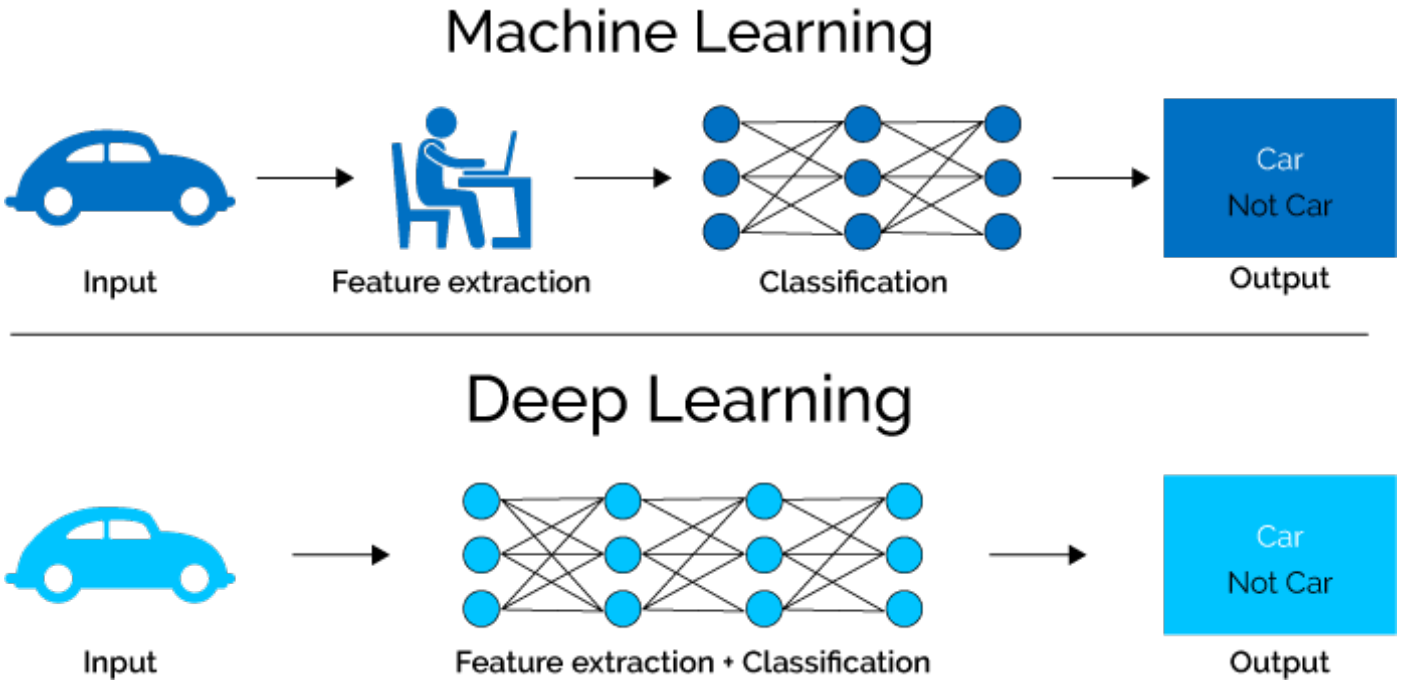
Die Analyse beginnt mit dem, was Mensch und Führung ausmacht: Sprache und Kommunikation. Der Vorstand und nach ihm die Führungskräfte der ersten Ebene experimentierten mit der neuen Technik. Manch einer berichtete von dem „mulmigen Gefühl", einen Computer anzurufen. Viele fanden es einfach nur cool. Fast alle waren skeptisch, doch die Neugierde lockte sie ans Telefon, für ein „Gespräch" mit künstlicher Intelligenz und mit der Aussicht, bei minimalem Zeitaufwand mehr über sich und das eigene Sprachverhalten herauszufinden. Führungskräfte müssen sich zunächst selbst verstehen, um in der Lage zu sein, andere zu führen.

Besser

Die Technologie zerlegt Sprachproben in Informationsbausteine und analysiert mit einem System von 500.000 Messpunkten linguistische und prosodische (akustische) Daten. Diese Daten werden anschließend zu psychologisch validen Aussagen über

Stulle, K. P. (Hrsg.) (2018), „Psychologische Diagnostik durch Sprachanalyse: Validierung der PRECIRE®-Technologie für die Personalarbeit", Springer-Verlag. Zur Kritik siehe Kanning, U.P. (2018), Fachbuch im Fokus, Wirtschaftspsychologie aktuellanalyse.html

Maschinenlernen (ML) vs. Tiefenlernen (Deep Learning, DL)



Quellen: Medium.com; Deepai.org

Microsoft trainiert Chatbot Tay 2016 durch den Austausch mit Twitter-Nutzern

NORDAKADEMIE GRADUATE SCHOOL

Nur wenige Stunden benötigte Tay, um sich von einem harmlosen Chatprogramm auf Twitter in einen frauenverachtenden Rassisten mit Sympathien für Adolf Hitler zu verwandeln.

Ähnliches Problem: Bei der Frage in Google „Juden sind“ wurde vom Algorithmus vorgeschlagen „böse“. Aufgrund von Millionen Aussagen im Internet, besonders in sozialen Medien, die das so sagen, wurde das durch DL so gelernt. Google konnte das nicht lösen, weil neuronale Netze in den tiefen Schichten unkontrollierbar sind.

Bots bilden In- und Out-Group

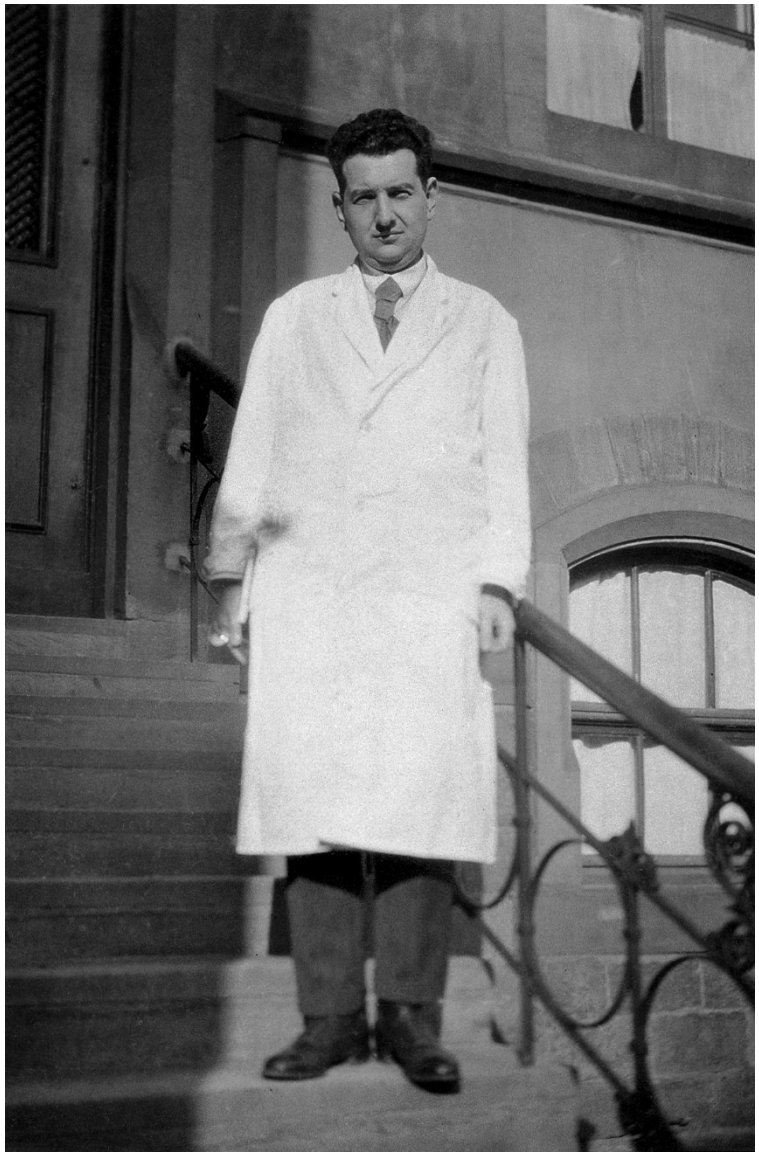
NORDAKADEMIE
GRADUATE SCHOOL

Die Forscher um den Informatiker Roger Whitaker ließen in Computersimulationen 100 intelligente Bots in Gruppen miteinander interagieren. Konkret ging es darum, in einem Geben-und-Nehmen-Spiel entweder einem Mitglied der eigenen Gruppe oder einem Bot aus einem anderen Team eine Belohnung zukommen zu lassen.

Im Laufe vieler tausend Simulationen zeigten die Bots eine zunehmend stärkere Sympathie für die eigene Gruppe: Wurde anfangs noch ohne erkennbare Bevorzugung belohnt, gingen Gruppenfremde mit fortschreitender Dauer des Experiments immer häufiger leer aus.

Gehorsamkeitsstudie nach Milgram (1963)

NORDAKADEMIE
GRADUATE SCHOOL



Gehorsamkeitsstudie nach Milgram (1963)

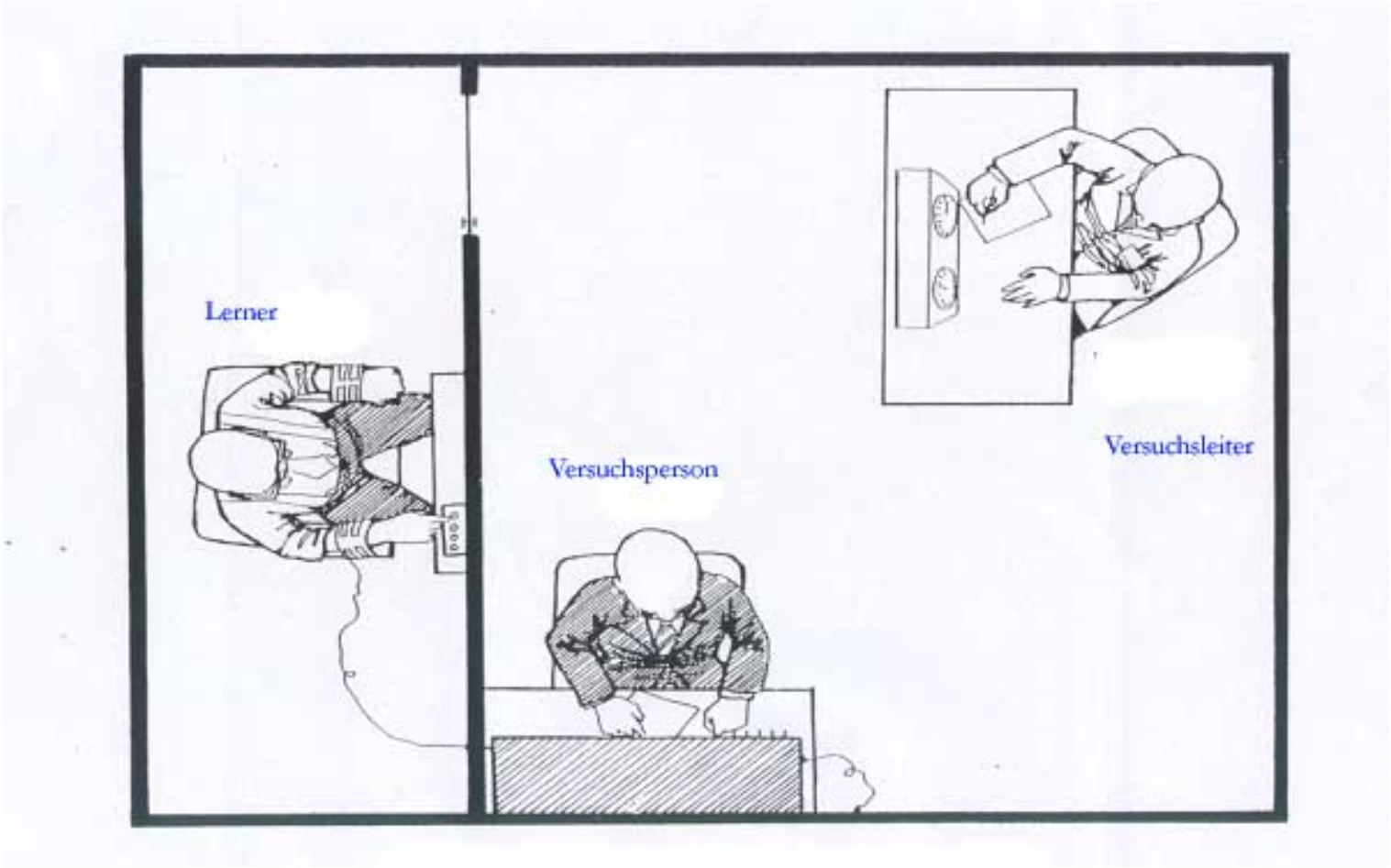
NORDAKADEMIE GRADUATE SCHOOL

Situation: Vp angeblich Assistent in Versuch zur Beziehung Lernleistung - Strafe

UV: Schrittweise erhöhte Stromschlagdosis bis in tödlichen Bereich

AV: Ausführung der Bestrafung

Aufbau Milgram-Studie



Rückmeldungen an Versuchsperson

NORDAKADEMIE GRADUATE SCHOOL

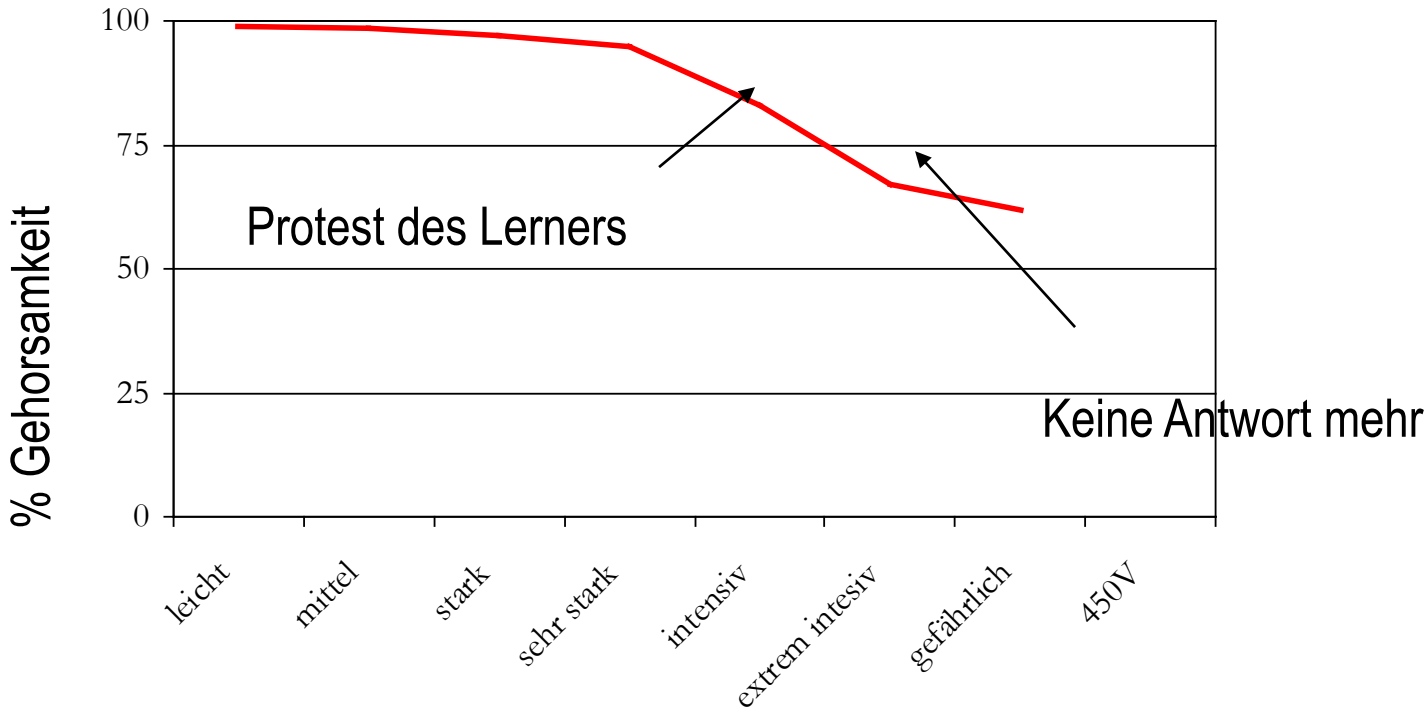
Lerner:

a. Beschwerde
b. Schmerzlaut
c. „Aufhören“ und Schmerzlaut
d. Lauter Schrei
e. „Gebe keine Antwort mehr“
f. Keine Reaktion

Versuchsleiter:

1. Fahren Sie fort
2. Das Experiment erfordert, daß Sie fortfahren.
3. Es ist absolut wesentlich, daß Sie fortfahren
4. Sie haben keine Wahl, fahren Sie fort

Ergebnisse Milgram-Studie



Der „Luzifer-Effekt“ (siehe Fiske, Harris & Cuddy sowie Zimbardo)

NORDAKADEMIE
GRADUATE SCHOOL

Handwerk für Despoten

1. Eine Ingroup-Outgroup-Unterscheidung erzeugen (Beispiel: die belgischen Kolonialherren in Ruanda „Hutu“ und „Tutsi“)
2. Durch Propaganda die Mitglieder der Outgroup als „Kakerlaken“, „Ratten“ oder „menschlicher Abschaum“ entmenschlichen (Dehumanisierung)
3. Gehorsamkeit gegenüber Autorität und Konformität mit Peers durchsetzen (Deindividuation) – das Individuum gibt jegliche Verantwortung an die Gruppe und dessen Führer ab
4. Zulassen, dass die Mitglieder der Out-Group wie Tiere behandelt werden, da sie ja „Unmenschen“ sind, Angehörige einer anderen Art, oder anonyme Nummern

Polarisierter Diskurs vor dem 1. Weltkrieg: z.B. Zar Nicholas II und Kaiser Wilhelm II (siehe Text-Analyse von David Winter)

NORDAKADEMIE
GRADUATE SCHOOL

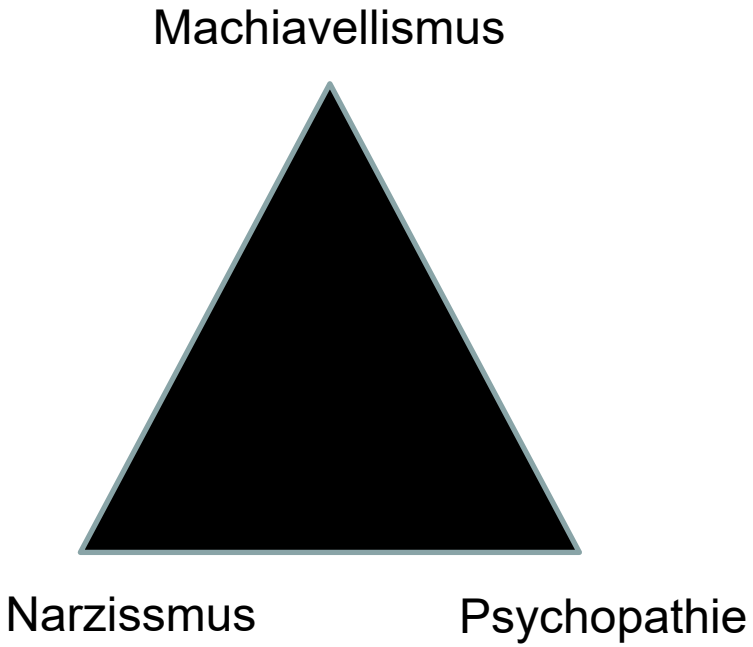
Pepper4Car

Unterstützung in der Pfleg

Ein anderer Erklärungsansatz …

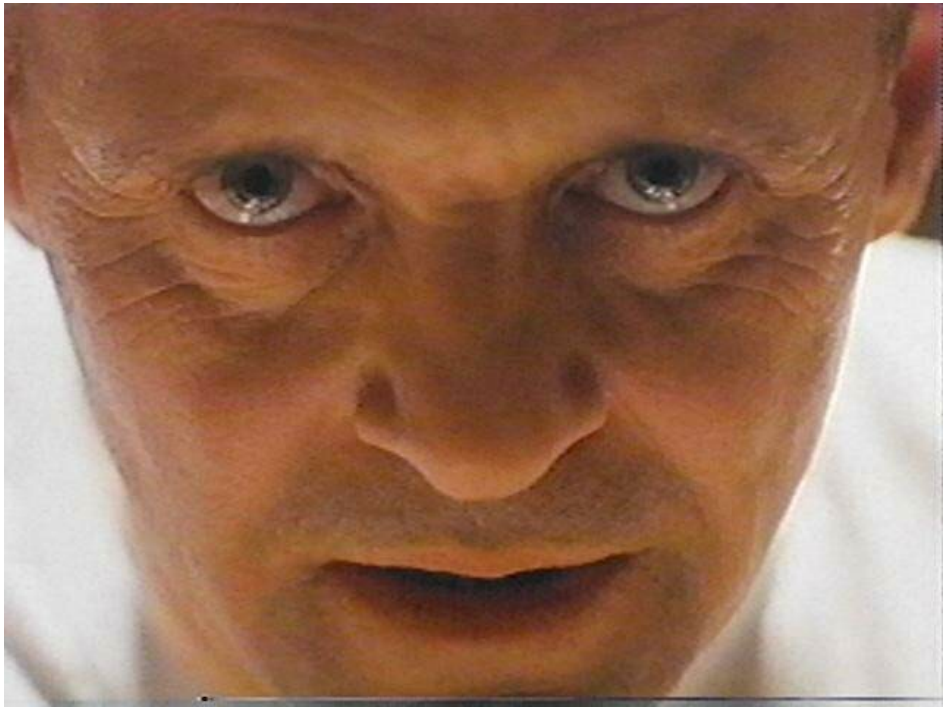
NORDAKADEMIE
GRADUATE SCHOOL

„Die dunkle Triade“



Psychopathie

- Psychopathen haben ein geringes Maß an Empathie
- Starke impulsive Neigungen
- Lügen, sind dabei aber „charmant“
- Wenig Ängstlichkeit, lernen nicht durch Strafe



Film
Dr. Hannibal Lecter
– gespielt von
Antony Hopkins –
in „Das Schweigen
der Lämmer“)

Siehe Robert Hare:
„Without Consciousness“

Siehe auch. „Der Fall Niels Högel – 106 Morde in der Pflege“

KI muss die Fragen auf die Einhaltung bestimmter Kriterien (Syntax) prüfen

NORDAKADEMIE
GRADUATE SCHOOL

- Generalisierung (Bezeichnungen von Gruppen, etc.)
- Tilgungen (fehlende Bezugsindizes, wie Person X ist „geeignet“ = im Vergleich zu wem, wer hat gemessen, was bedeutet „geeignet“, wie steht es um die zeitliche Gültigkeit, etc.)
- Falsche Prozessbeschreibungen („sie macht mich ärgerlich“ integriert „sie“ in emotionale Bewertung
- Nominalisierungen (ein Prozess, wie sich ärgern, wird zu einem Ding „Ärger“ – eine Liste typischer Keywords kann einfach erstellt werden)
- Universal-Aussagen (zu prüfen ist auf Worte wie „alle, jeder, immer“ etc.)

Herzlichen Dank!

NORDAKADEMIE
GRADUATE SCHOOL

In der Kommunikation mit dem Nutzer kann das System

NORDAKADEMIE GRADUATE SCHOOL

- mit vertiefenden Fragen zu einer Klärung auffordern (welche Person jüdischen Glaubens meinen Sie konkret?)
- eine gegebene Aufgabe korrekt umformulieren („Der Mitarbeiter Meier macht Ärger“ in „Sie ärgern sich über den Mitarbeiter Meier?“)
- zeigen, welche Parameter für eine korrekte Beantwortung fehlen

Systemimmanente Konflikte der menschlichen Selbststeuerung

NORDAKADEMIE
GRADUATE SCHOOL

- Durchsetzung (das Recht auf Entscheidungsfreiheit und eigenverantwortlichem Handeln) versus Sicherheit (das Recht auf körperliche, seelische und geistige Unversehrtheit)
- Integration (das Recht auf Zugehörigkeit und Teilhabe) versus Individualität (das Recht aus Anerkennung und darauf, sein Leben nach eigener Fasson zu gestalten)
- Rationalität (das Recht auf freien Zugang zu allen Wissensquellen und sachlich begründeter Distanz) versus Emotionalität (das Recht auf Offenheit, Einfühlung, emotionaler Nähe)

KI entspricht „Unbewussten“ – sie generiert Ergebnisse aus Input

NORDAKADEMIE
GRADUATE SCHOOL

- Zeitliche oder räumliche Blockaden
- Gefahren für Körper, Seele, Geist
- Vereinsamung (Kündigung, Trennung)
- Verachtung
- Informationen vorenthalten
- Emotionale Ausgrenzung

Der polarisierte Diskurs in den Sozialen Medien erhält sich von alleine Aufrecht

NORDAKADEMIE GRADUATE SCHOOL

- Bei Posts die ein hohes Machtmotiv aufweisen, lässt sich auch in den Kommentaren ein hohes Machtmotiv finden. Insbesondere bei den Kommentaren die am häufigsten geliked wurden
- Social Media Posts mit einem hohen Machtmotiv werden häufiger geteilt bzw. es wir auf diese mehr reagiert in der Anzahl von Likes, Kommentaren usw.
- Sicherheit als Defizitwert: ein Wert, dessen Bedeutung für uns nach oben eskaliert, wenn wir ihn für zu niedrig erachten – und der uns gleichzeitig alle anderen Werte vergessen lässt

→ Suche nach „starken Männer“, die uns retten

Narzissmus

NORDAKADEMIE
GRADUATE SCHOOL

Narzissten

- finden sich großartig, einzigartig, kontrollieren gerne und benötigen steten „Applaus“ (Dammann, 2007)
- übertreiben eigene Leistungen, blocken Kritik ab, wirken arrogant und kompromisslos (O’Boyle et al., 2012)
- nehmen gerne Führungspositionen ein und
- werten andere Personen ab (Keller Hansbrough & Jones, 2014)
- …

Narzissus von Carravagio;
http://commons.wikimedia.org/wiki/File:Michelangelo_Caravaggio_065.jpg

Machiavellismus

NORDAKADEMIE
GRADUATE SCHOOL

- Neigung unmoralisch und opportunistisch zu handeln
- Andere Menschen werden manipuliert, um die eigenen Ziele zu erreichen („Empathie mit Tunnelblick“)
- Verhält sich sozial gewünscht, solange dies seinen Zielen entspricht.

Machiavelli, Detail aus einem Bild von Santi di Tito