Verantwortung und Erklärbarkeit von Künstlicher Intelligenz

Warum KI verständlich sein muss

Impulsvortrag

05. Februar 2018

Timo Speith, B.A., M.Sc.
Universität des Saarlandes,
Theoretische Philosophie

Verantwortung und Erklärbarkeit von Künstlicher Intelligenz

Warum KI verständlich sein muss

Impulsvortrag

05. Februar 2018

Timo Speith, B.A., M.Sc.
Universität des Saarlandes,
Theoretische Philosophie

Weil es sonst inakzeptable Verantwortungslücken gibt!

Was hat KI mit Verantwortung zu tun?

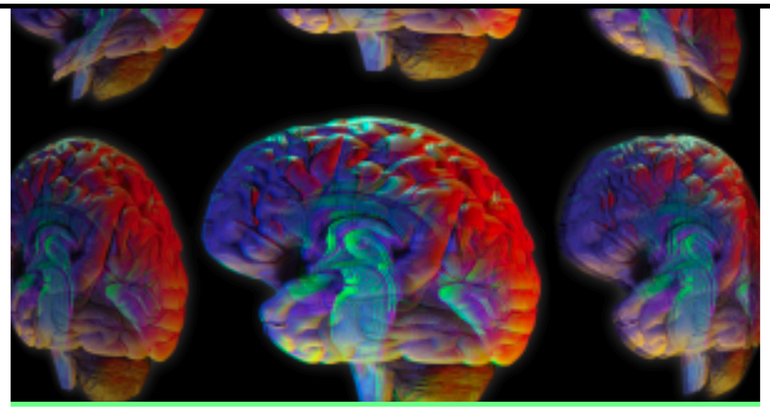
Wo finden wir KI?

22. Oktober 2018, 05:48 Uhr Digitalisierung

Arbeit aus dem Automaten

https://www.sueddeutsche.de/digital/digitalisierung-arbeitslosigkeit-jobcenter-1.4178635

/ Futurism
/ The Byte



SCI-FI VISIONS

Should Coma Patients Live or Die? Machine Learning Will Help Decide.

An algorithm is helping Chinese researchers determine if a coma patient will wake up again.

Victor Tangermann | September 27th 2018

https://futurism.com/machine-learning-coma-patients-live

Machine Bias

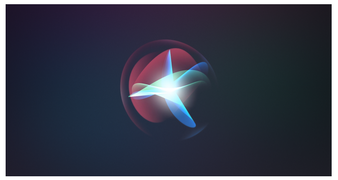
There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Folgenschwere und Verantwortung



Google

amazon

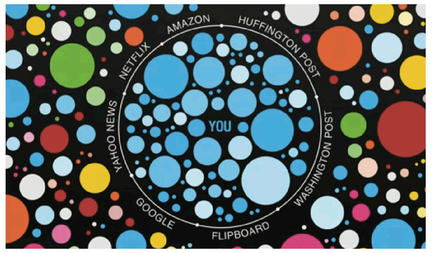
Musikauswahl

Suchanfragen zu Banalitäten

Einkaufsempfehlungen bei Amazon

Saugroboter

https://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles



Vorauswahl von News und Posts in sozialen Netzwerken

Visumvergabe

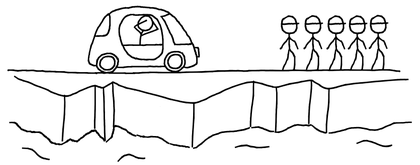
Kreditwürdigkeit

Strafzumessung

Abschalten lebensnotwendiger Geräte

Jobvergabe und Bewertung Arbeitssuchender

Autonome Fahrzeuge in Dilemmasituationen

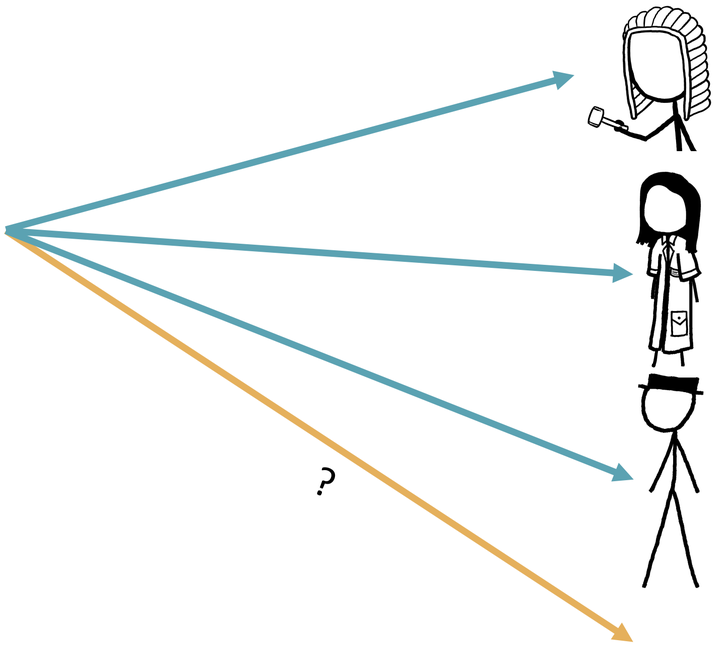


›Folgentrivial‹ ⟷ Folgenschwer

Die **Frage der Verantwortung** stellt sich insbesondere in Fällen, in denen folgenschwere Entscheidungen getroffen werden müssen.

Folgenschwere und Verantwortung

In vielen solcher Fälle haben wir *Experten* und oft auch Zeit, sie miteinzubeziehen…



Strafzumessung

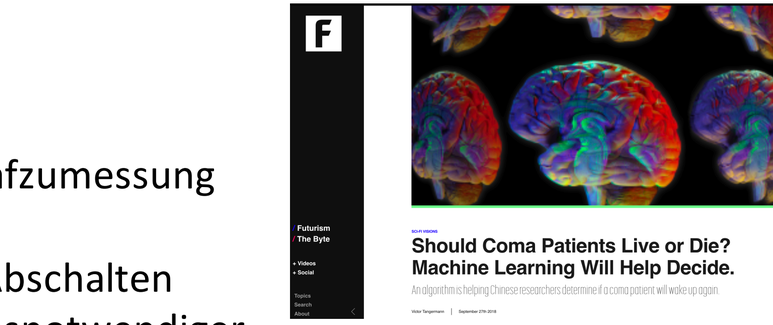
Abschalten lebensnotwendiger Geräte

Jobvergabe und Bewertung Arbeitssuchender

Autonome Fahrzeuge in Dilemmasituationen

Standardreaktion: *Nicht ersetzen, unterstützen*!

Allokation von Verantwortung scheint dann klar: Bei den Experten.


Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Futurism

The Byte

Should Coma Patients Live or Die? Machine Learning Will Help Decide.

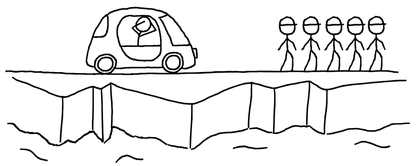
An algorithm is helping Chinese researchers determine if a coma patient will wake up again

Süddeutsche Zeitung

SZ.de Zeitung Magazin

22. Oktober 2018, 05:48 Uhr Digitalisierung

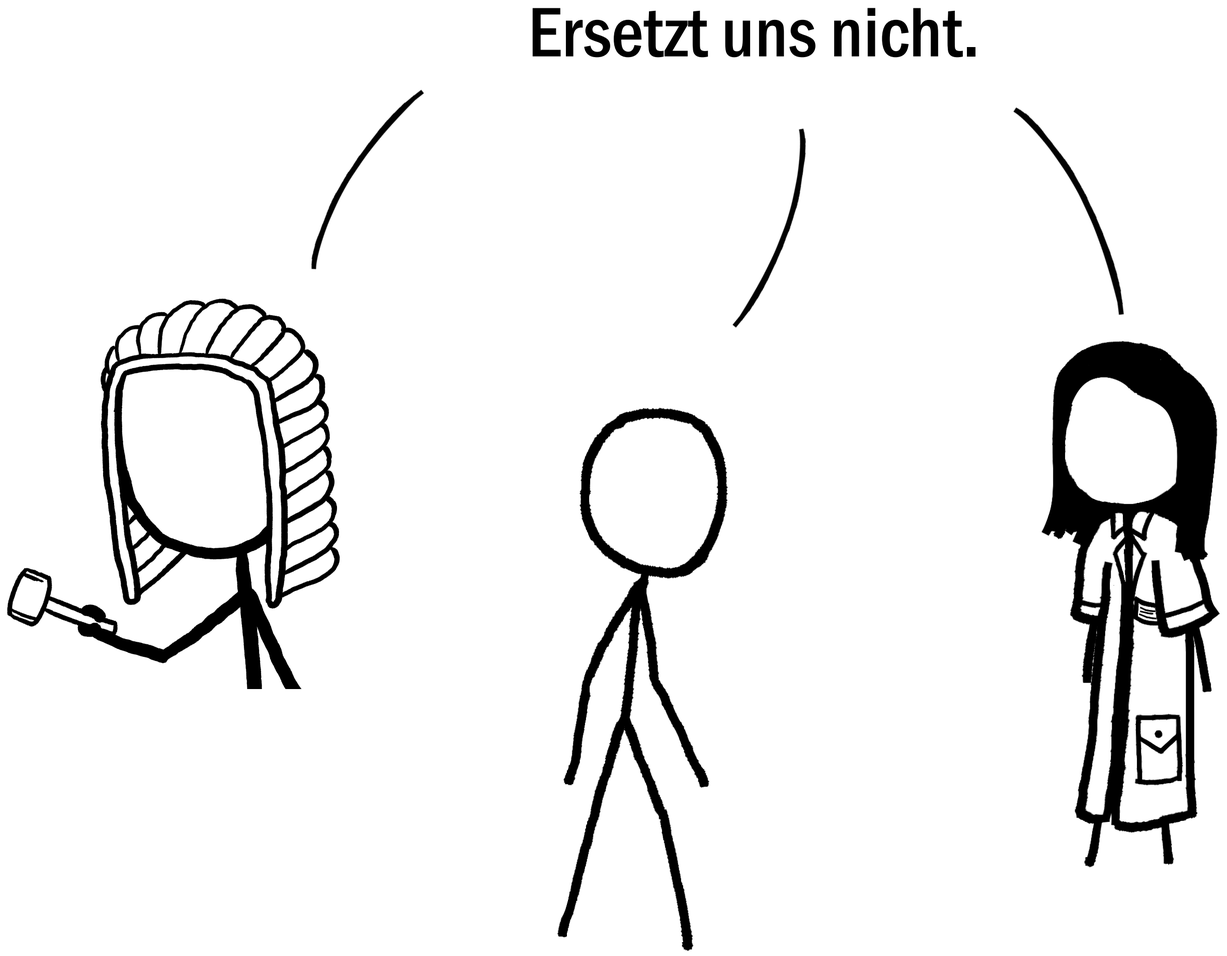
Arbeit aus dem Automaten



›Folgentrivial‹ ⟷ Folgenschwer

Die **Frage der Verantwortung** stellt sich insbesondere in Fällen, in denen folgenschwere Entscheidungen getroffen werden müssen.

Fazit 1

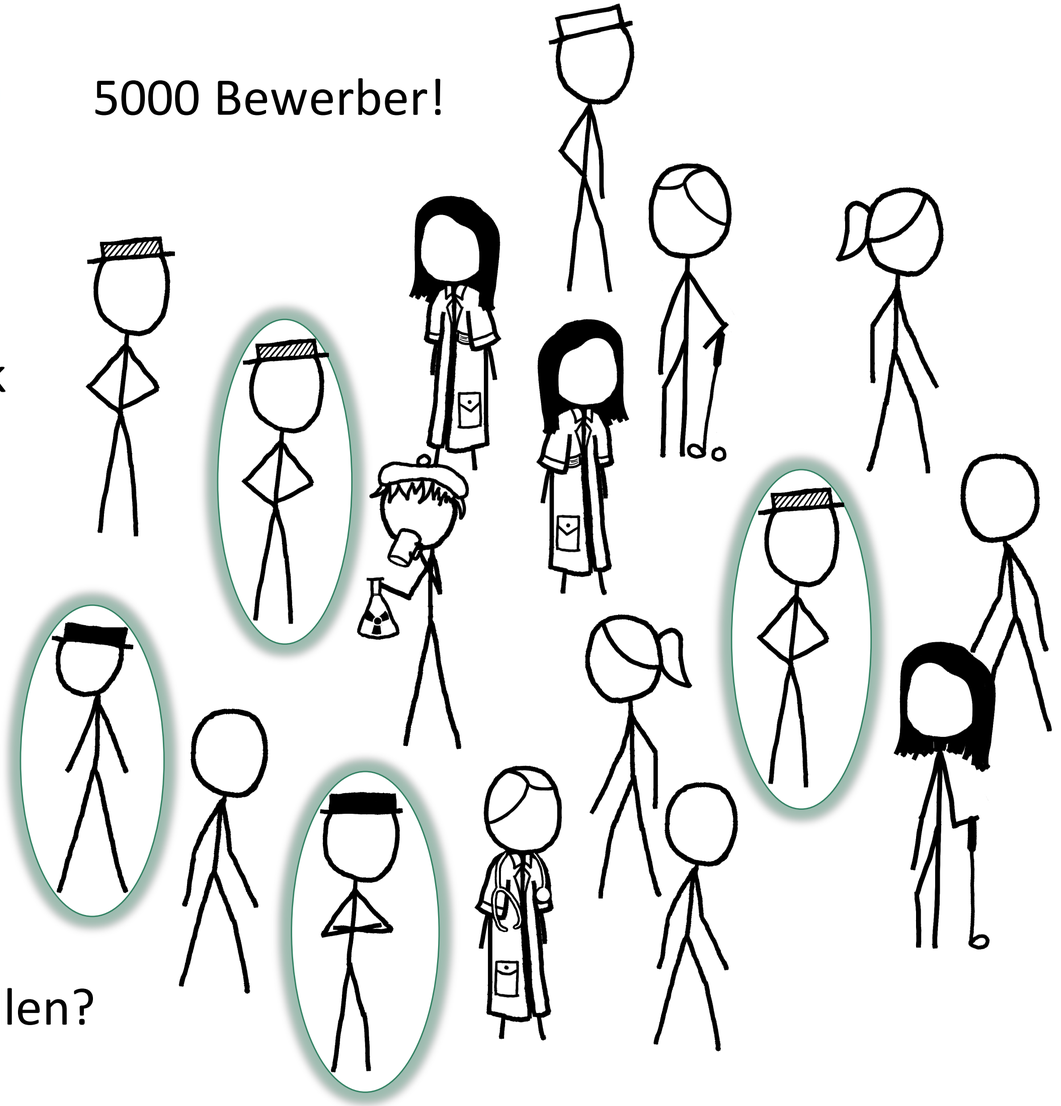


Sollten wir deshalb bei *folgenschweren* Entscheidungen gänzlich auf KI verzichten?

Gründe für Unterstützung: Überforderung und schlechte Heuristiken

- Das ist Rudolph, der Personaler. Er hat einen Job zu besetzen.

- 5000 Bewerber!



Eine solche Heuristik ist weder fair noch wünschenswert.

Kann KI helfen?

Wen soll Rudolph einstellen?

Gründe für Unterstützung: Vorurteile

BUSINESS NEWS OCTOBER 10, 2018 / 3:12 AM / 2 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

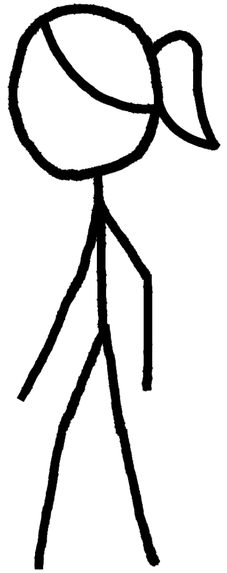
Rudolph, der rassistische Personaler

Das ist Rudolph, der Personaler, der auch Rassist ist.
Er hat einen Job zu besetzen.

Dieser Zustand ist inakzeptabel.

Kann KI helfen?

Das ist Susanne, die sich auf den Job beworben hat.
Sie ist am besten qualifiziert.
Allerdings ist sie auch von anderer Ethnie.



Wer bekommt den Job wohl nicht?

Gründe für Unterstützung: Kognitive Verzerrungen

Lunch Luck

Published online 11 April 2011 | Nature | doi:10.1038/news.2011.227

News

Hungry judges dispense rough justice

When they need a break, decision-makers gravitate towards the easy option.

Zoë Corbyn

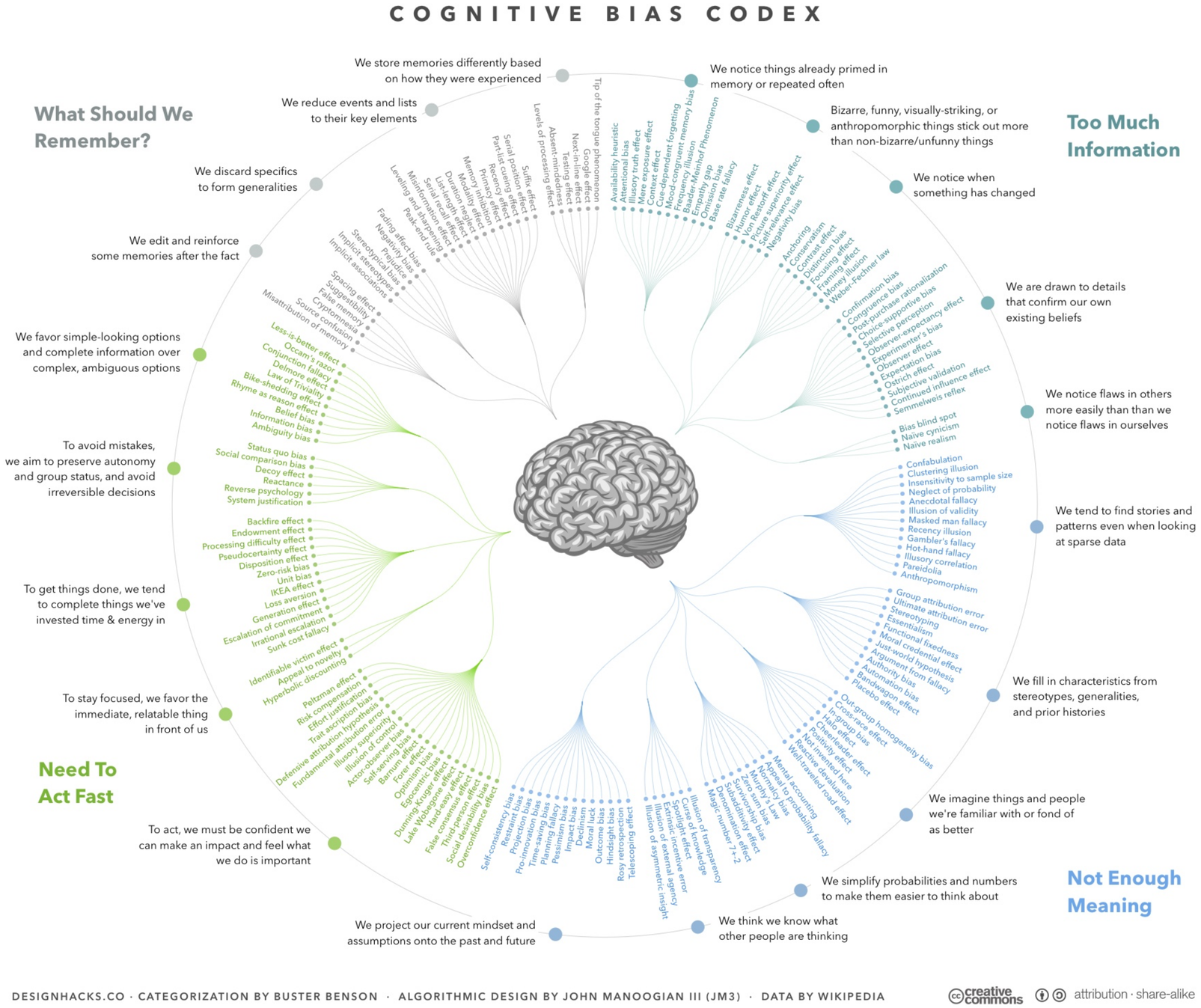
A prisoner's chance of parole depends on when the judge hearing the case last took a break, say researchers who have studied decisions in Israeli courts. As judges tire and get hungry, they slip towards the easy option of denying parole, say the researchers.

The bias could apply in any situation in which people

A judge is less likely to parole a prisoner at the end of a session than at the beginning.

Punchstock

http://www.nature.com/news/2011/110411/full/news.2011.227.html



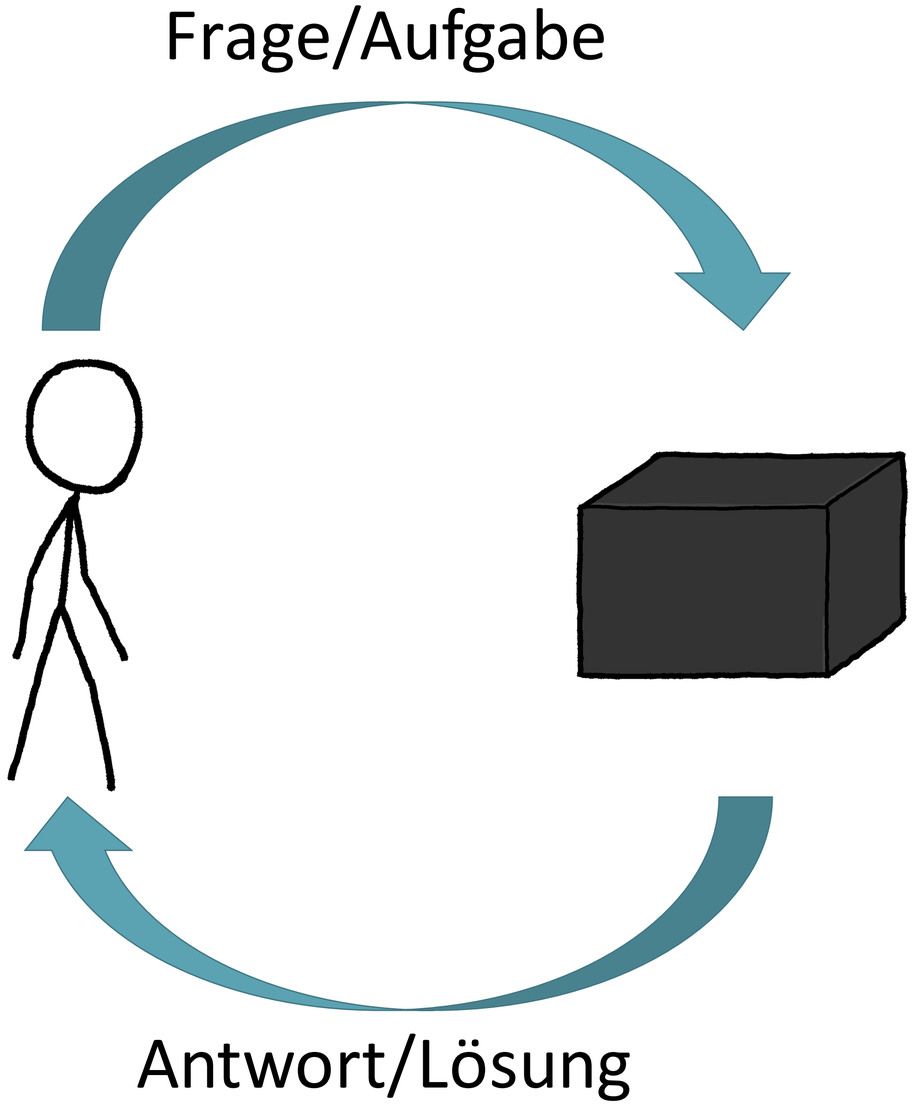
https://en.wikipedia.org/wiki/List_of_cognitive_biases
(recht lang, wenig überraschend...)

In vielen Gebieten der vielversprechendste Ansatz

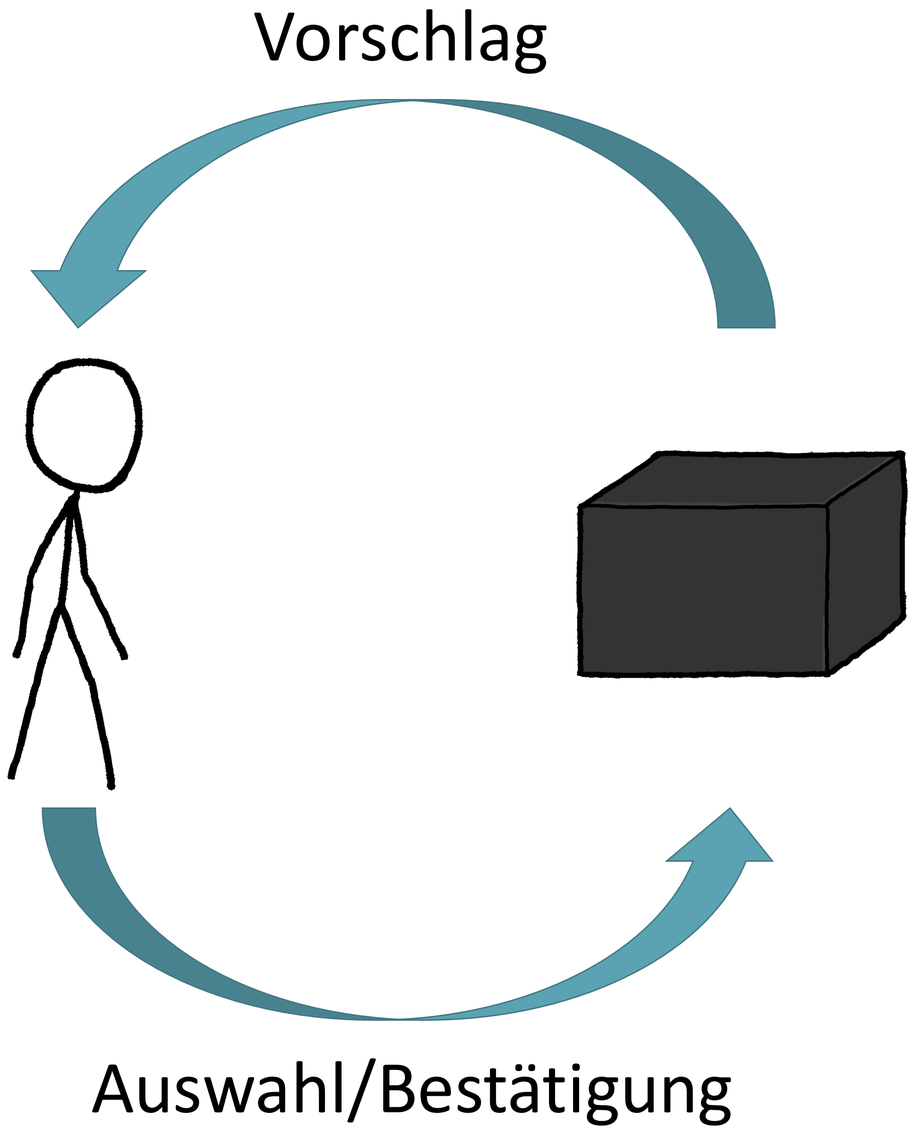
- Wir nutzen beides:
  - Die Möglichkeiten der KI
    - Prädiktive Kraft/Reduzierung von Unsicherheit
    - ›Unbestechlichkeit‹
    - Systematik
  - Die Fähigkeiten der Menschen
    - Können hinterfragen
    - ›Sehen‹ u.U. mehr

    (z.B. Gründe und Kausalität statt bloße Korrelation, schwer bis unmöglich messbare Eigenschaften, Einzelfallspezifisches)

Human in the Loop: Pull

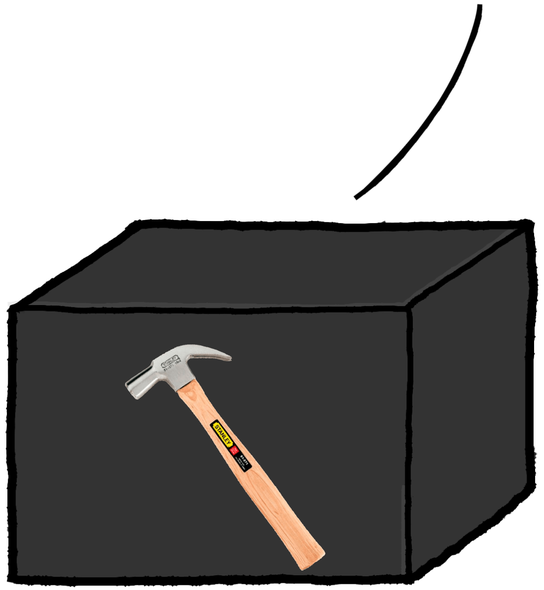


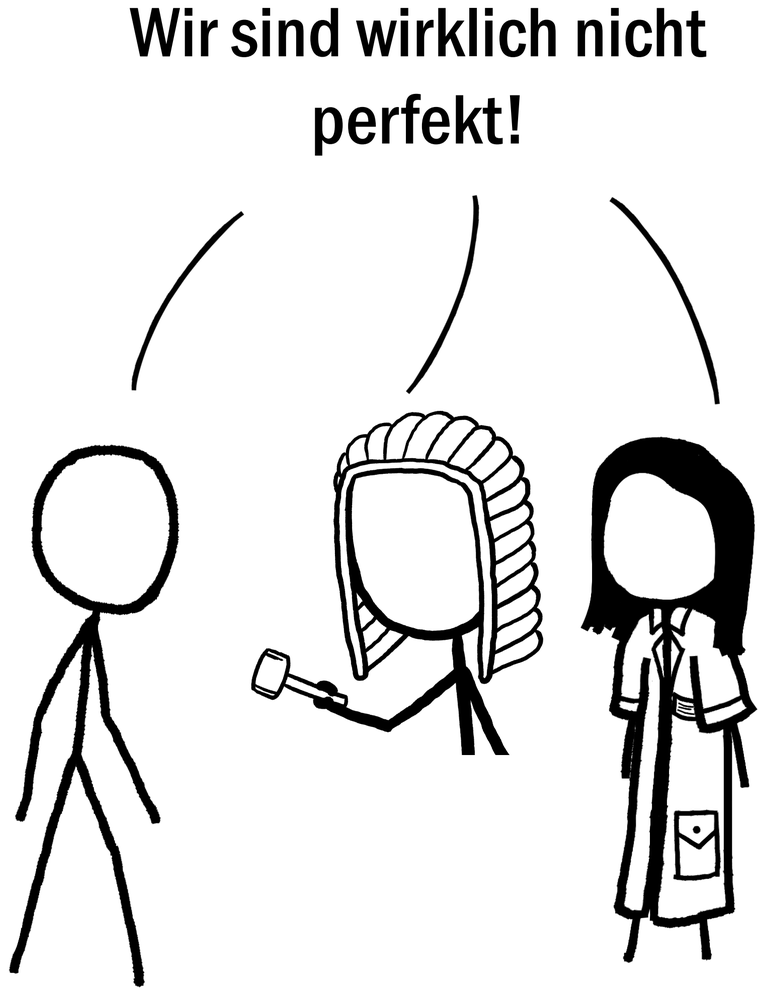
Human in the Loop: Push



Fazit 2

Zusammen sind wir besser.





Warum es so einfach nicht ist:

Das Recht auf Anfechtbarkeit und ›eingebaute Werte‹

Beispiel COMPAS

https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/
https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html
https://en.wikipedia.org/wiki/Loomis_v._Wisconsin

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)

- Proprietärer, closed-source Risikobewertungsalgorithmus (kein ML), der in vielen amerikanischen Jurisdiktionen zum Einsatz kommt.
- Konkreter Fall:
  - Eric L. Loomis wurde zu sechs Jahren Gefängnis verurteilt.
  - Der Algorithmus ›urteilte‹: »a high risk of violence, high risk of recidivism, high pretrial risk.«

https://www.wsj.com/articles/wisconsin-supreme-court-to-rule-on-predictive-algorithms-used-in-sentencing-1465119008

Beispiel COMPAS

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)

- Loomis versuchte das Urteil anzufechten, dies wurde aber vom zuständigen Wisconsin Supreme Court abgelehnt:

¶6 The court of appeals certified the specific question of whether the use of a COMPAS risk assessment at sentencing "violates a defendant's right to due process, either because the proprietary nature of COMPAS prevents defendants from challenging the COMPAS assessment's scientific validity, or because COMPAS assessments take gender into account."[12]

¶28 In denying the post-conviction motion, the circuit court explained that it used the COMPAS risk assessment to corroborate its findings and that it would have imposed the same sentence regardless of whether it considered the COMPAS risk scores. Loomis appealed and the court of appeals certified the appeal to this court.

¶56 Additionally, this is not a situation in which portions of a PSI are considered by the circuit court, but not released to the defendant. The circuit court and Loomis had access to the same copy of the risk assessment. Loomis had an opportunity to challenge his risk scores by arguing that other factors or information demonstrate their inaccuracy.

http://www.scotusblog.com/wp-content/uploads/2017/02/16-6387-op-bel-wis.pdf

Beispiel COMPAS

https://assets.documentcloud.org/documents/2702103/Sample-Risk-Assessment-COMPAS-CORE.pdf

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)

32. If you lived with both parents and they later separated, how old were you at the time?
☑ Less than 5 ☐ 5 to 10 ☐ 11 to 14 ☐ 15 or older ☐ Does Not Apply

33. Was your father (or father figure who principally raised you) ever arrested, that you know of?
☑ No ☐ Yes

34. Was your mother (or mother figure who principally raised you) ever arrested, that you know of?
☑ No ☐ Yes

Please think of your friends and the people you hung out with in the past few (3-6) months.

39. How many of your friends/acquaintances have ever been arrested?
☐ None ☐ Few ☑ Half ☐ Most

64. Do you have an alias (do you sometimes call yourself by another name)?
☑ No ☐ Yes

Beispiel COMPAS

https://assets.documentcloud.org/documents/2702103/Sample-Risk-Assessment-COMPAS-CORE.pdf

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)

78. How strongly do you agree or disagree with the following: I always behaved myself in school?
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

95. How often did you feel bored?
☐ Never ☑ Several times/mo ☐ Several times/wk ☐ Daily

Criminal Personality

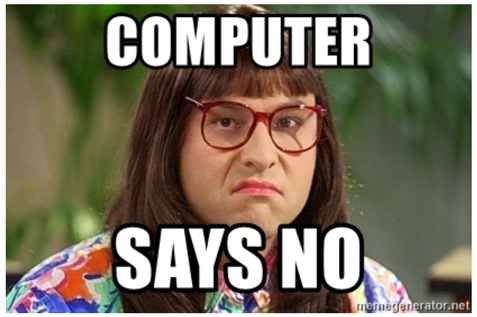
The next few statements are about what you are like as a person, what your thoughts are, and how other people see you. There are no 'right or wrong' answers. Just indicate how much you agree or disagree with each statement.

112. "I am seen by others as cold and unfeeling."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

137. "Some people just don't deserve any respect and should be treated like animals."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

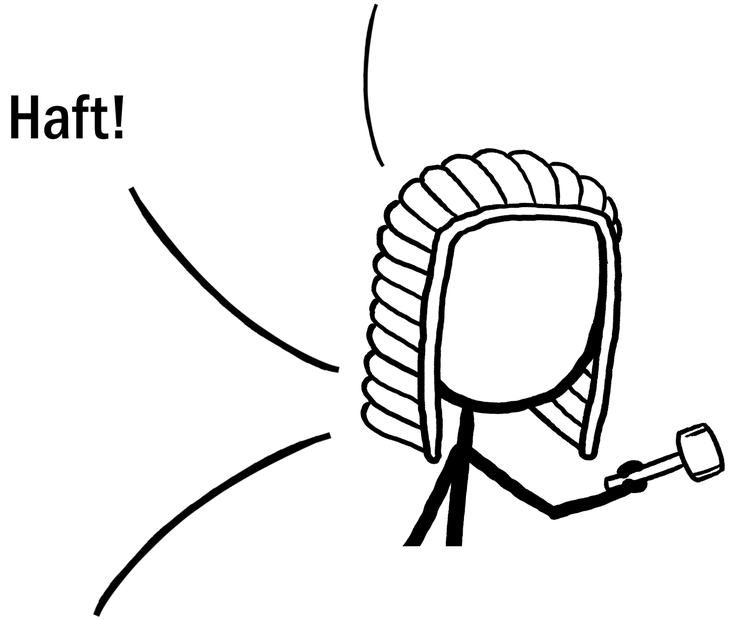
Eine verantwortungsvolle Entscheidung?

Darum.



Darum.

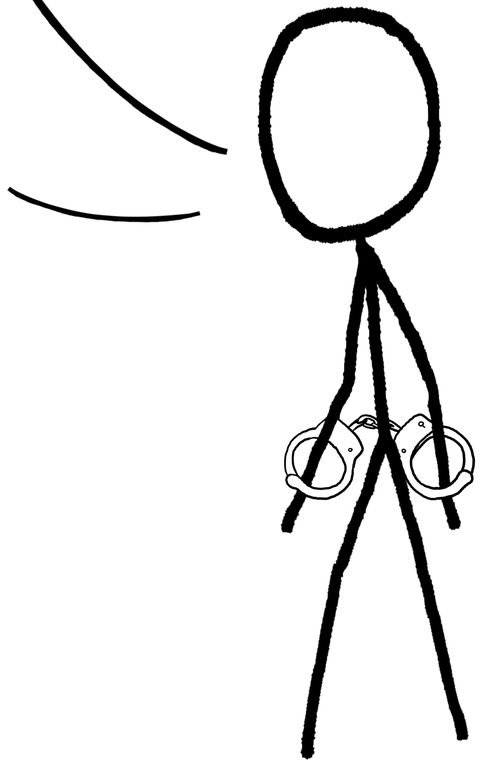
Haft!



Okay, klar, Ihr gutes Recht. Justitia, warum noch gleich?

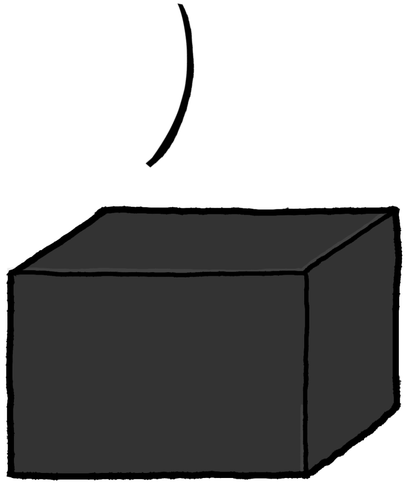
Da wüsste ich aber gerne, wieso!

Inakzeptabel!

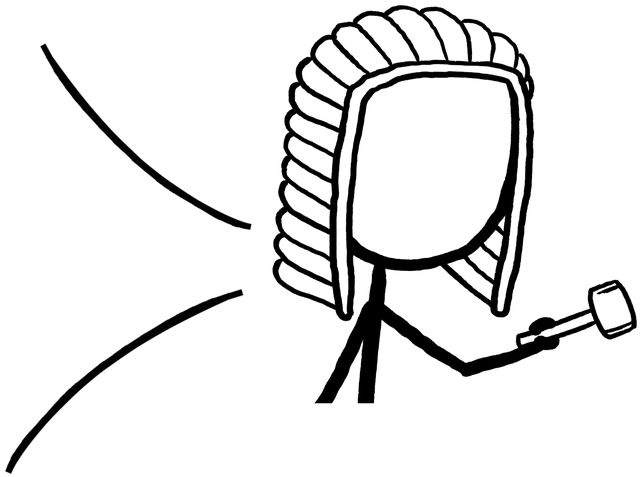


Eine verantwortungsvolle Entscheidung?

Klar doch.

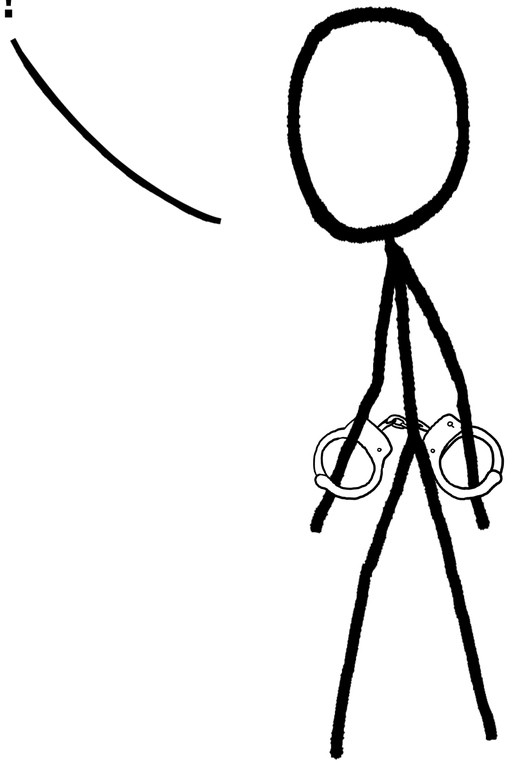


Haft!



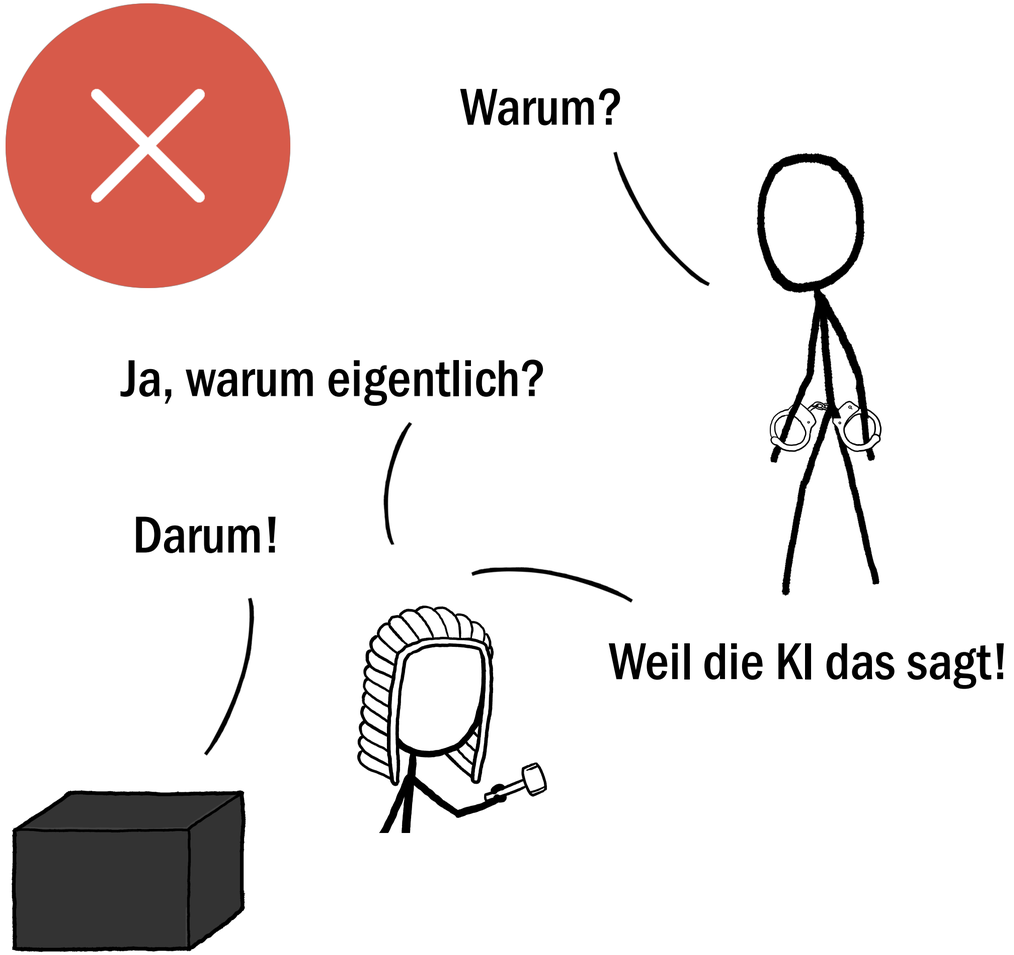
Okay, klar, Ihr gutes Recht. Justitia, schicke bitte meine Urteilsbegründung inklusive Deiner begründeten Einschätzung an seinen Anwalt.

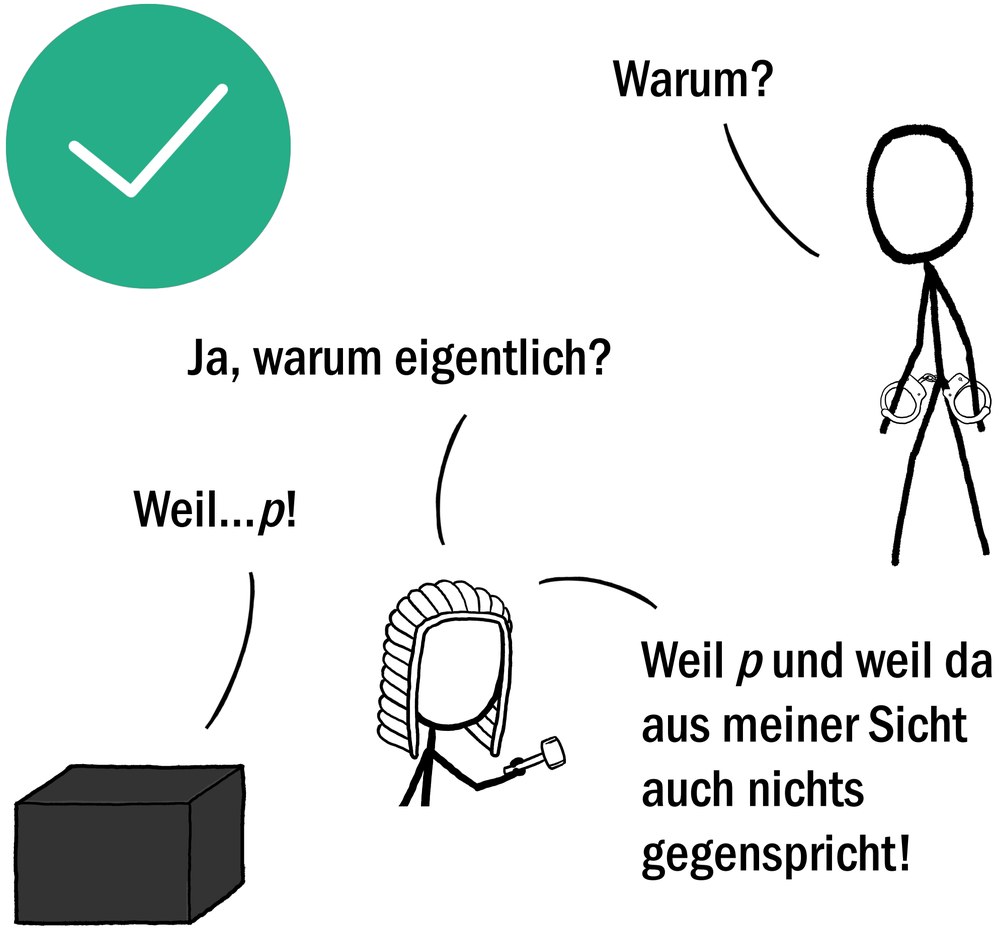
Da wüsste ich aber gerne, wieso!



Fazit 3

In einer liberalen Demokratie, in einem Rechtsstaat, geht es nicht ohne Erklärbarkeit.



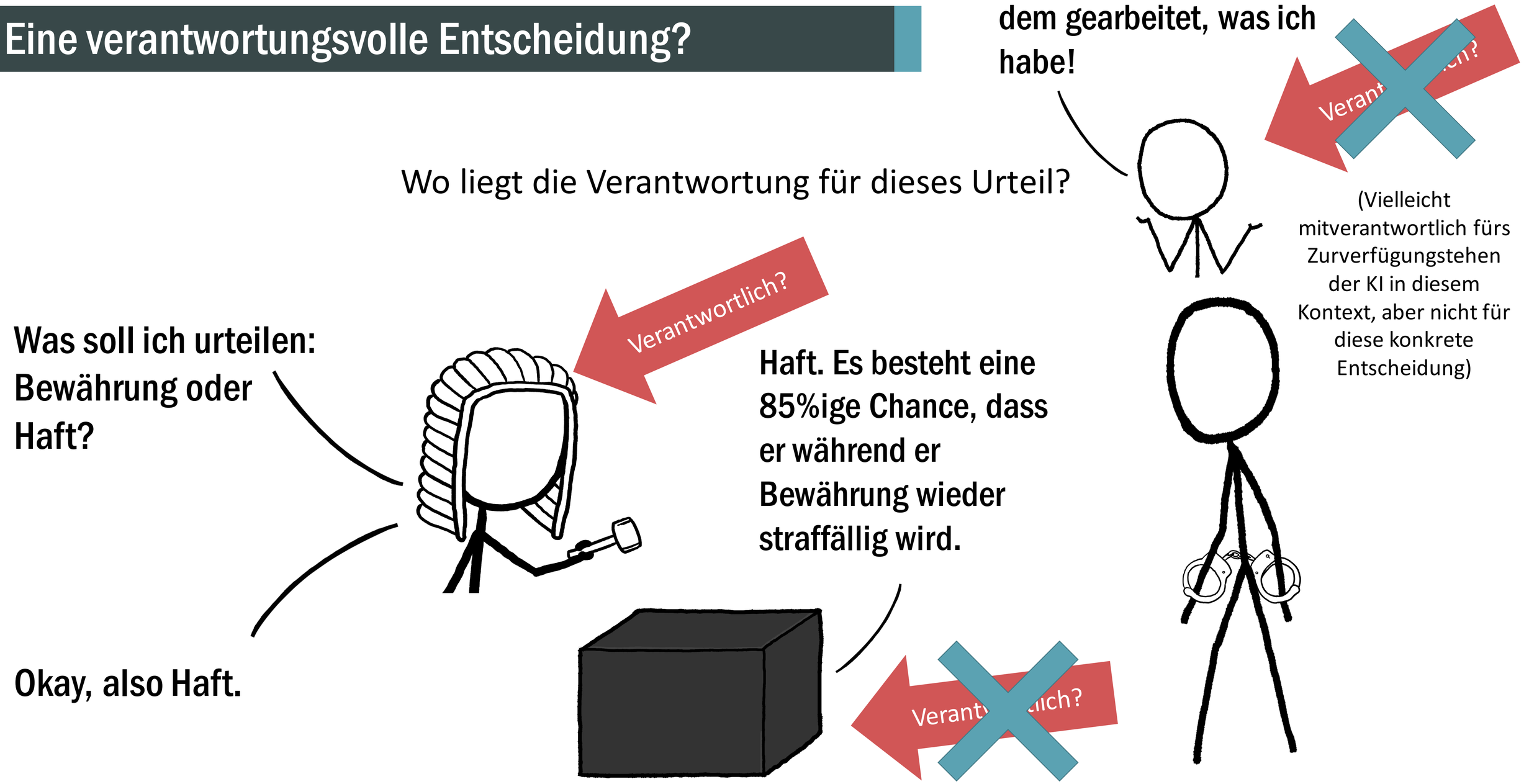


Und was hat das jetzt mit Verantwortung zu tun?

Wir haben doch beides: Verantwortung und KI-Power!

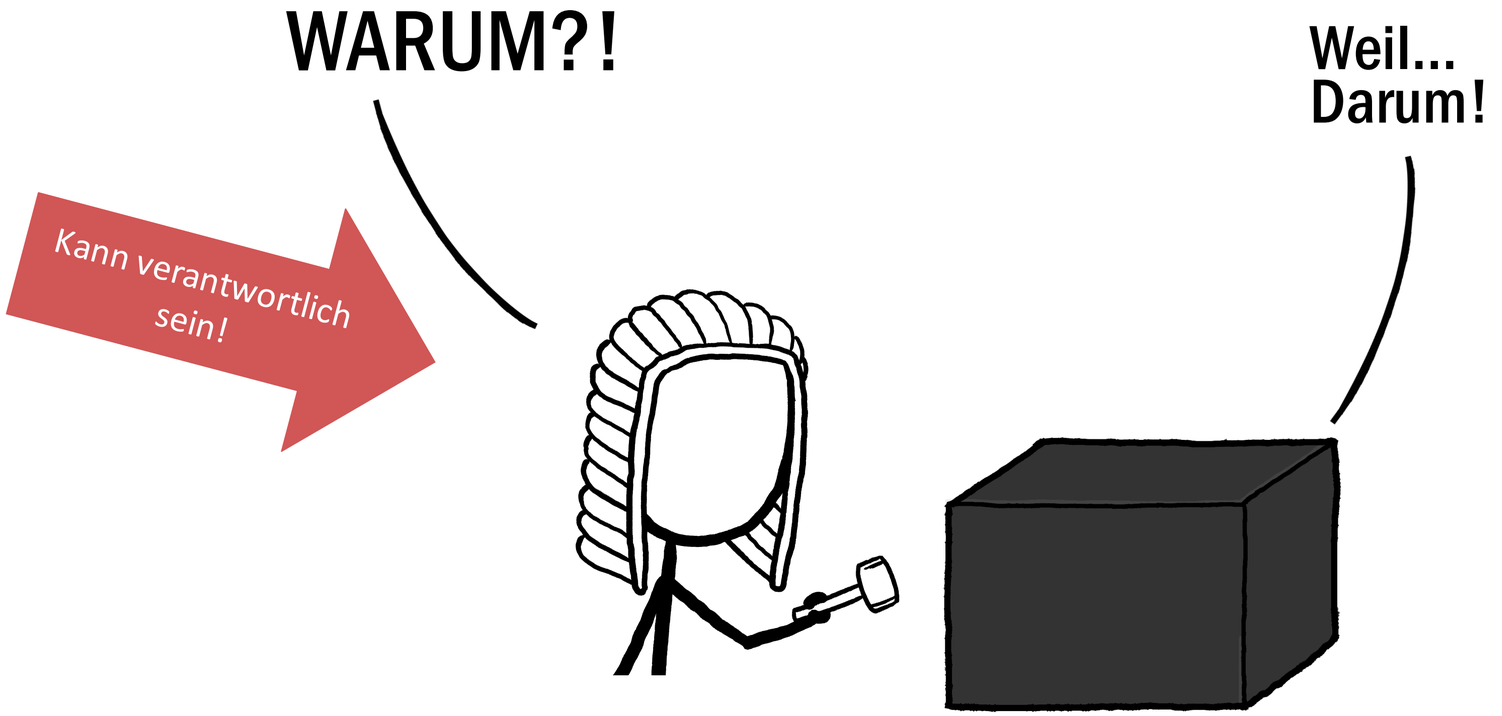
Eine verantwortungsvolle Entscheidung?

Wo liegt die Verantwortung für dieses Urteil?



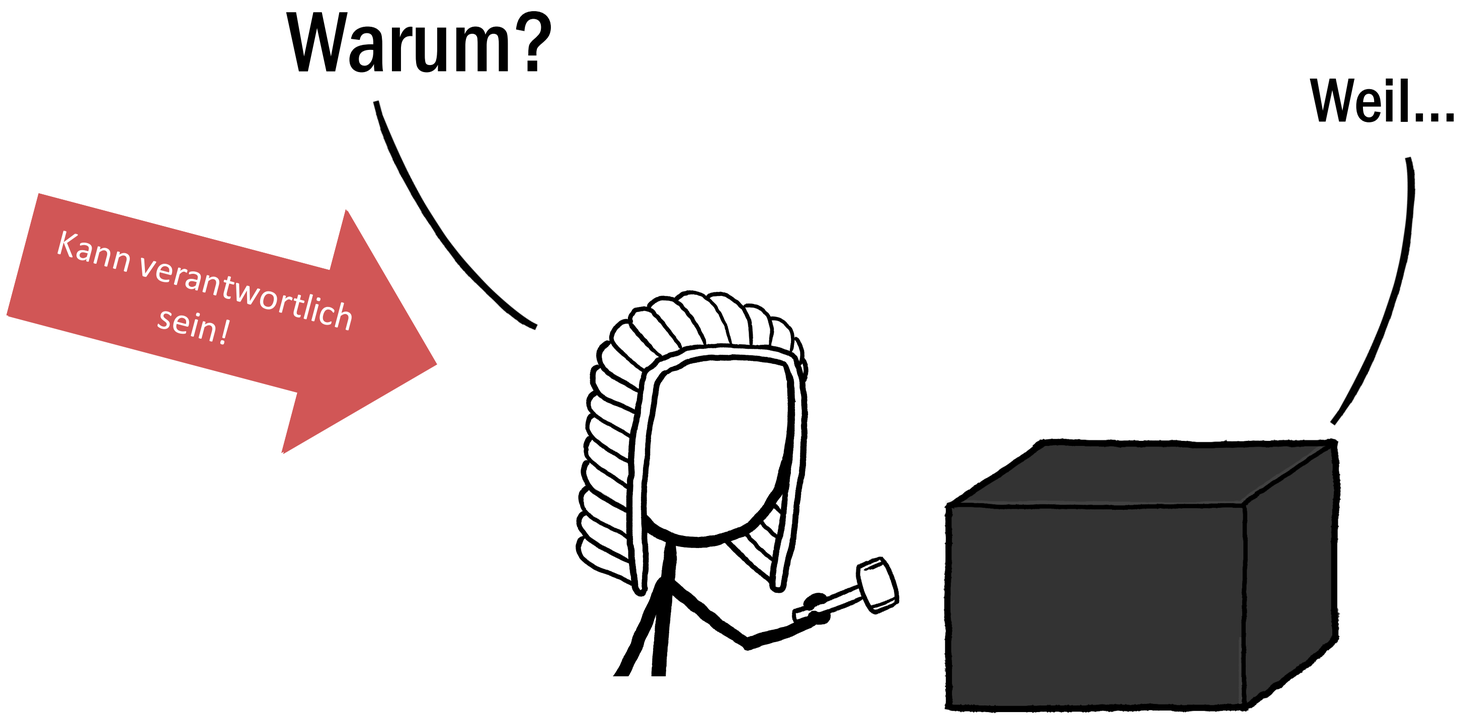
Ein Lösungsansatz

- Ein Mensch kann nur dann im Lichte einer Empfehlung verantwortungsvoll entscheiden, wenn er die Grundlage jener Empfehlung gegen seine eigene Grundlage abwägen kann.
- Nur dann kann er sich bewusst nach Abwägung von Für und Wider *für* oder *gegen* das Befolgen der Empfehlung entscheiden.



Ein Lösungsansatz: Erklärbarkeit

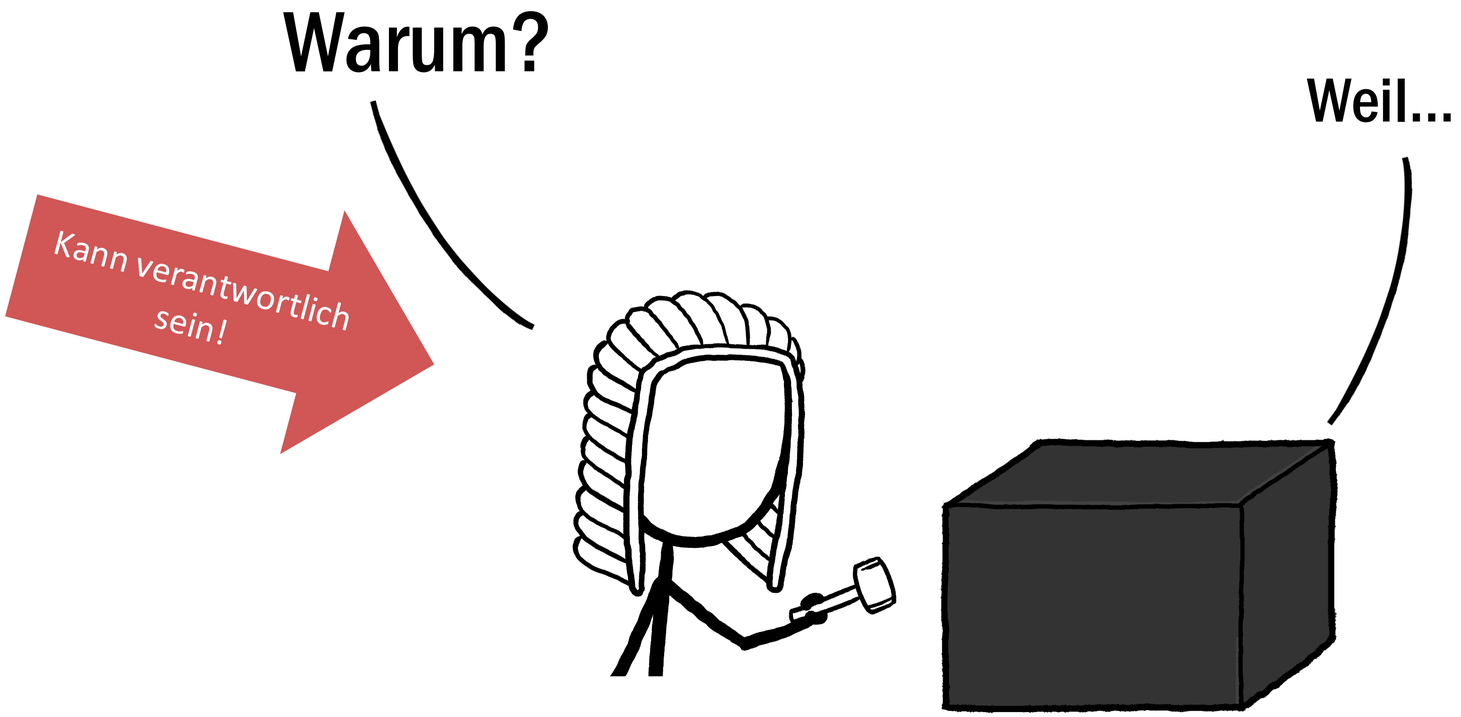
- Es muss mindestens auf Anfrage *Erklärungen* zu Empfehlungen geben, d.h. dargelegte Gründe.



KI muss *verständlich*, *hinterfragbar* und *rechtfertigbar* sein.

Fazit 4

Erklärbarkeit ist *notwendige Voraussetzung* für verantwortungsvolle, KI-gestützte Entscheidungen.



Wie können wir Erklärbarkeit gewährleisten?

Wir wissen es *noch* nicht …

… aber wir arbeiten dran!

Aktuelles
Uni-Porträt
Fakultäten|Einrichtungen
Wirtschaftsportal

https://www.uni-saarland.de/nc/universitaet/aktuell/artikel/nr/19923.html

Montag, 26. November 2018

Neuer Sonderforschungsbereich: Softwaresysteme sollen ihr Verhalten selbst erklären

Selbst Experten verstehen das Verhalten komplexer Softwaresysteme immer weniger. Dabei regeln diese inzwischen immer stärker unseren Alltag, sei es als intelligente Haussteuerung, im autonomen Fahrzeug oder in der industriellen Produktion. Wissenschaftler der Universität des Saarlandes, zweier Max-Planck-Institute und der Technischen Universität Dresden wollen jetzt in einem neuen Sonderforschungsbereich Mechanismen entwickeln, die nicht nur Experten, sondern auch Laien das Verhalten komplexer Softwaresysteme besser vermittelt. Die Deutsche Forschungsgemeinschaft fördert dieses Großprojekt mit elf Millionen Euro über vier Jahre hinweg.

FOUNDATIONS OF PERSPICUOUS SOFTWARE SYSTEMS

— Enabling Comprehension in a Cyber-Physical World —

https://www.saarbruecker-zeitung.de/sz-spezial/hochschule/neuer-sonderforschungsbereich-an-der-saar-uni_aid-34732051

später lesen

26. November 2018 | 17:28 Uhr

Saarländische Spitzenforschung

Diese Programme erklären sich selbst

Holger Hermanns ist der Sprecher des neuen Sonderforschungsbereichs an der Universität des Saarlandes. Die Informatiker entwickeln gemeinsam mit der Technischen Universität Dresden selbstlernende Software. FOTO: Oliver Dietze

Teilen Twittern

Saarbrücken. Neuer Sonderforschungsbereich der Informatik an der Saar-Uni entwickelt die Software der Zukunft. **Von Peter Bylda** ›

Zwei Abschließende Botschaften

- Erreichen wir Erklärbarkeit nicht droht ein Novum:

  Zugewinn an technischem Vermögen
  → *prinzipieller* **Verlust an Erkenntnis und Kontrolle**

- Einsicht in die Grundlage einer Entscheidung ermöglicht erst **Vertrauenswürdigkeit** und **gerechtfertigte Akzeptanz**.

Danke für Ihre Aufmerksamkeit!