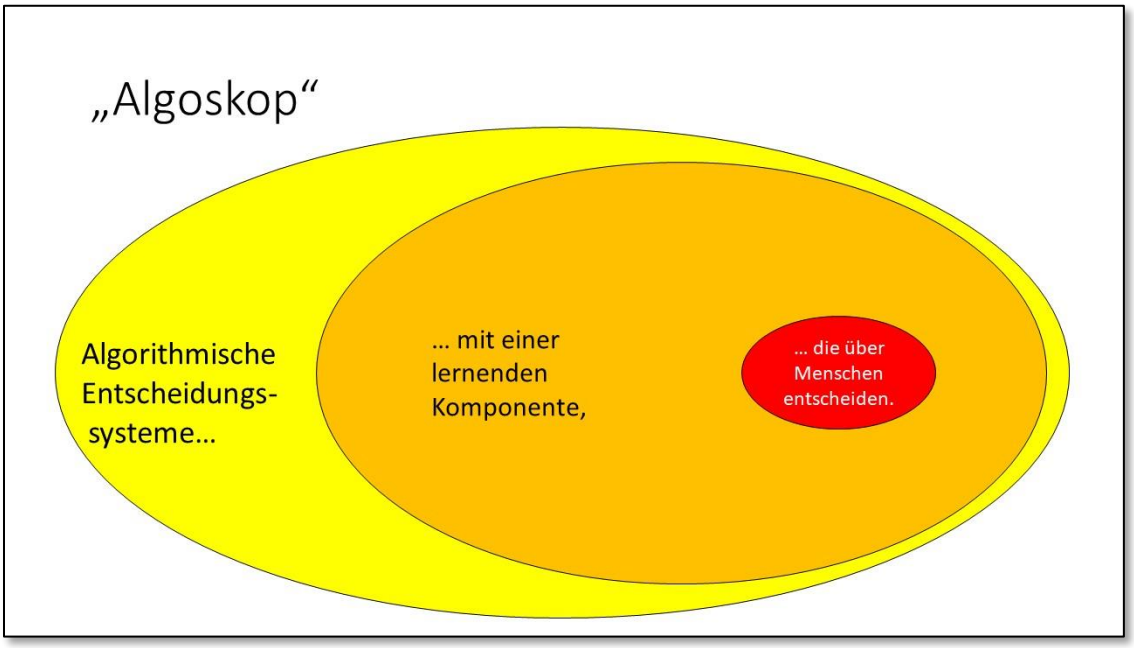
Handout für den Vortrag

"Black Box Analysen zur Kontrolle von ADM-Systemen"

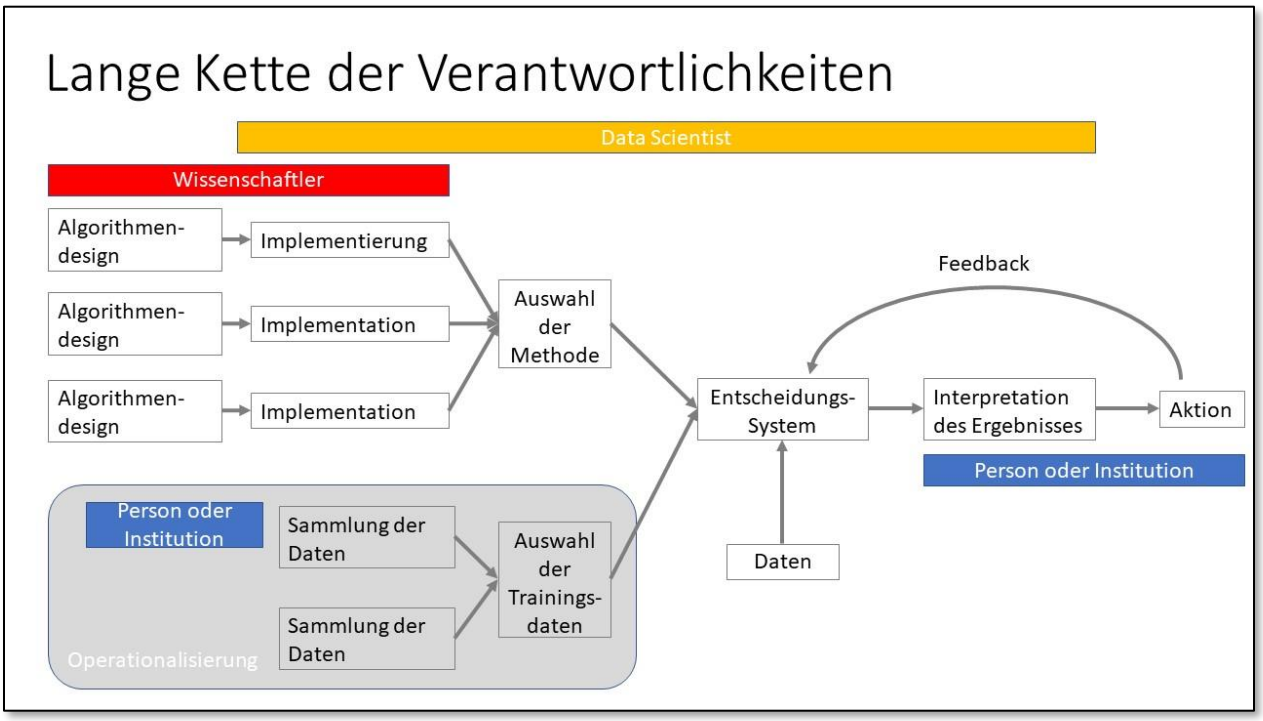
von Prof. Dr. Katharina A. Zweig,
Algorithm Accountability Lab, TU Kaiserslautern
zweig@cs.uni-kl.de, Tel: 0631-205-3346
Twitter: @nettwerkerin

1 Zusammenfassung Vortrag vom 14.10.2018 (1. Klausur der Enquete-Kommission KI)

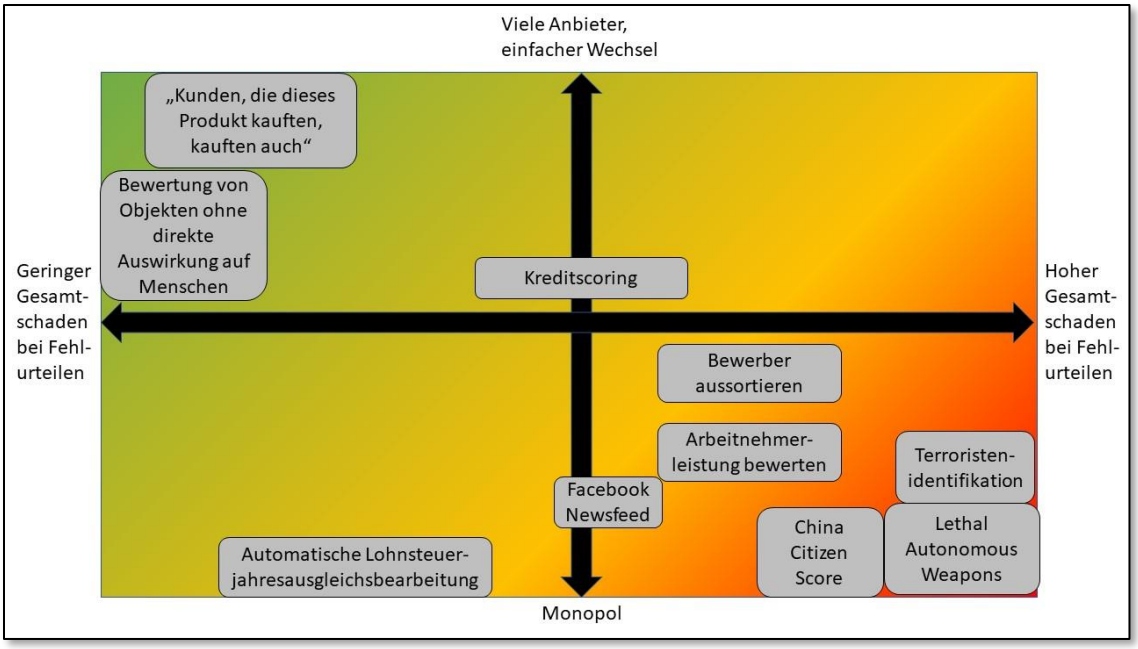


Wie schon am 14.10.2019 in der ersten Klausur der Enquete-Kommission KI berichtet, muss nur eine kleine Menge an algorithmen Entscheidungssystemen (*algorithmic decision making systems – ADM Systeme)* auf **technischer Ebene** reguliert werden. Das sind solche, die über Menschen oder Menschen nahestehende Objekte urteilen – insbesondere dann, wenn sie eine lernende Komponente beinhalten. Es ist **wichtig**, dass nur die kleinste Menge an ADM-

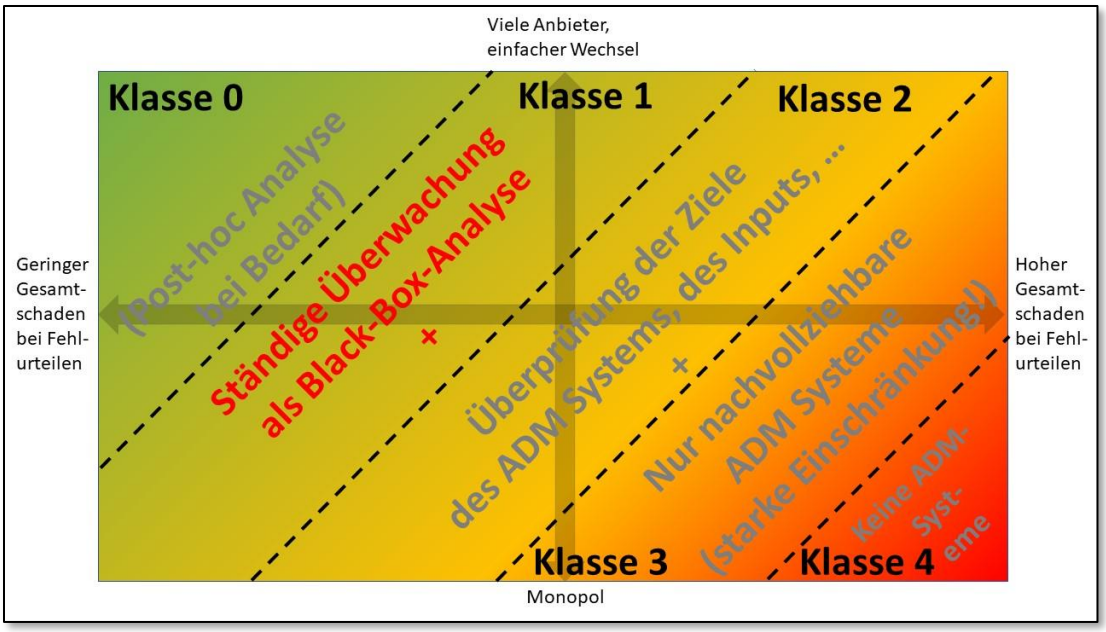
Systemen auf technischer Ebene reguliert wird, um keine **Innovationshemmnisse** aufzubauen.



Ebenfalls im letzten Vortrag wurde auch schon darauf hingewiesen, **warum** es so viele fehlerhafte ADM Systeme gibt, insbesondere dort, wo schwer definierbar soziale Konzepte wie beispielsweise "Vertrauen", "soziale Situation" oder "gebildet" vorhergesagt werden sollen. Das liegt vor allen Dingen daran, dass in der Entwicklung solcher Systeme sehr viele, verschiedene Personen beteiligt sind und die Kommunikation zwischen diesen Personen noch nicht etabliert ist. Die Lange Kette der Verantwortlichkeiten zeigt wer was tut, um mit Hilfe eines ADM-Systems eine Entscheidung zu treffen. Mehr dazu findet sich in der Studie: "Wo Maschinen irren können" (Zweig, Lischka & Fischer, 2018, Bertelsmann Stiftung, Reihe "Algoethik).

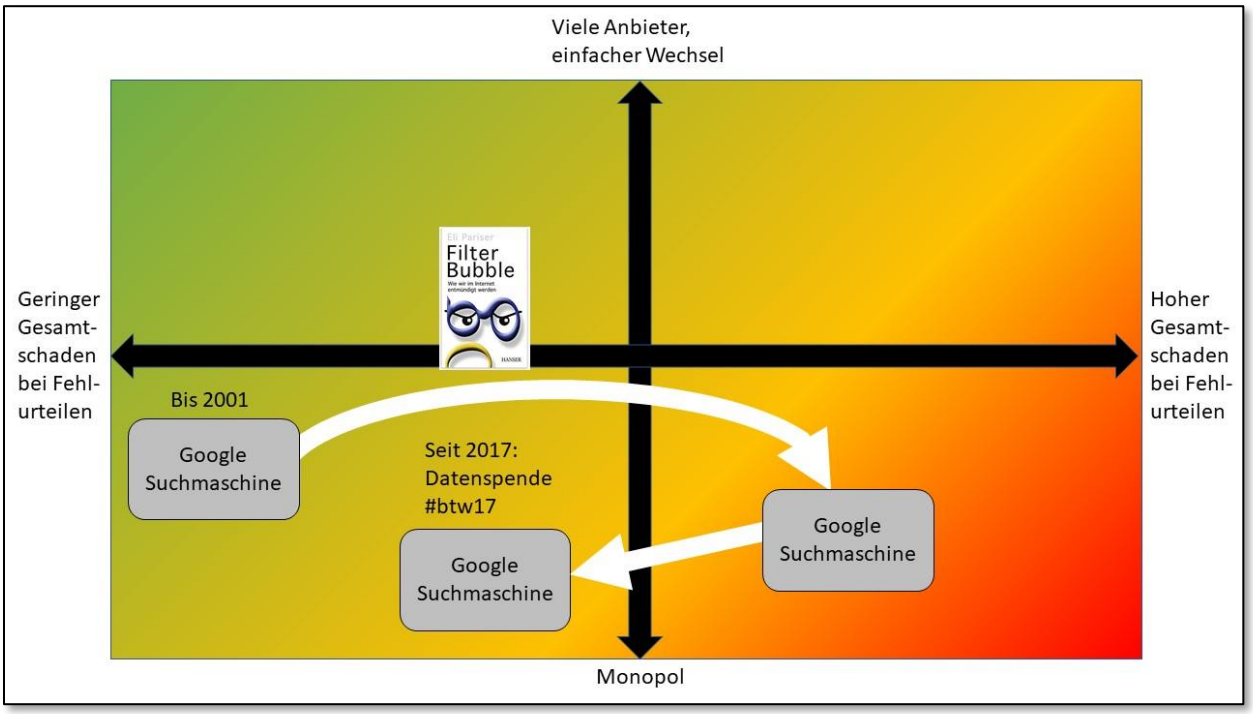


Ebenfalls vorgestellt wurde eine zweidimensionale Risikomatrix, bei der ADM-Systeme aufgrund ihres Gesamtschadenpotenzials und der Leichtigkeit, mit der eine Zweitmeinung eingeholt werden kann bzw. Widerspruch eingelegt werden, verortet werden.

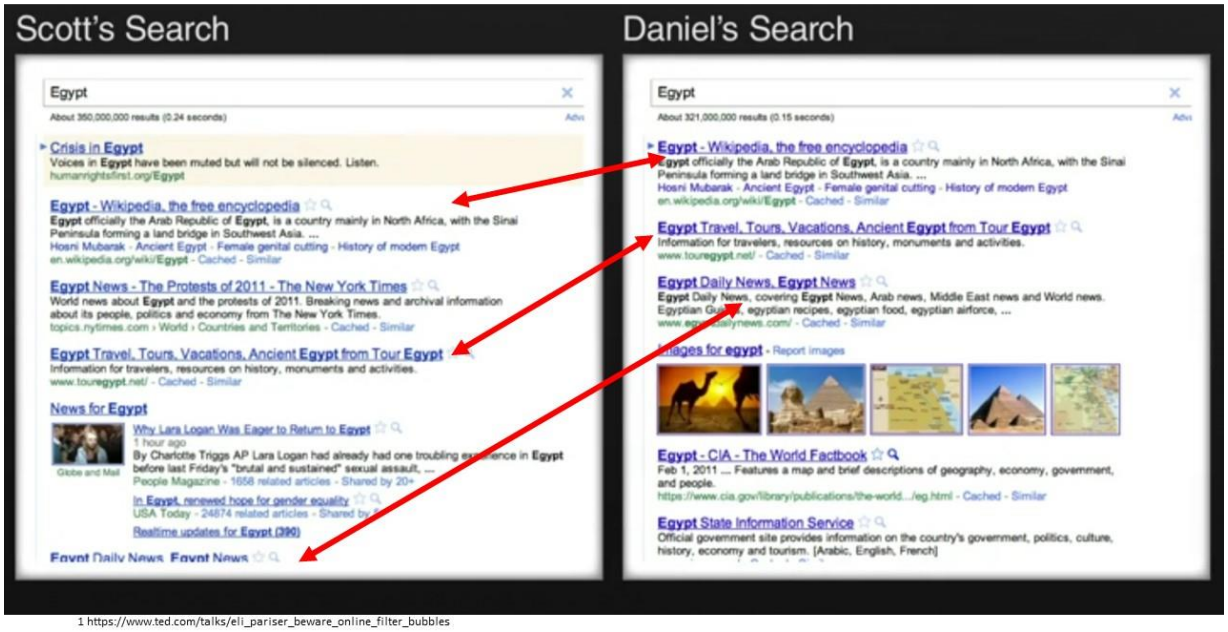


Die Verortung ergibt eine Risikoklasse, die mit verschiedenen Forderungen an die technische Regulierung verknüpft ist. In Klasse 0 gibt es keine Forderungen, ab Klasse 1 halten wir "Black-Box-Analysen" für notwendig.

2 Verortung von Googles Suchmaschine ändert sich über die Zeit

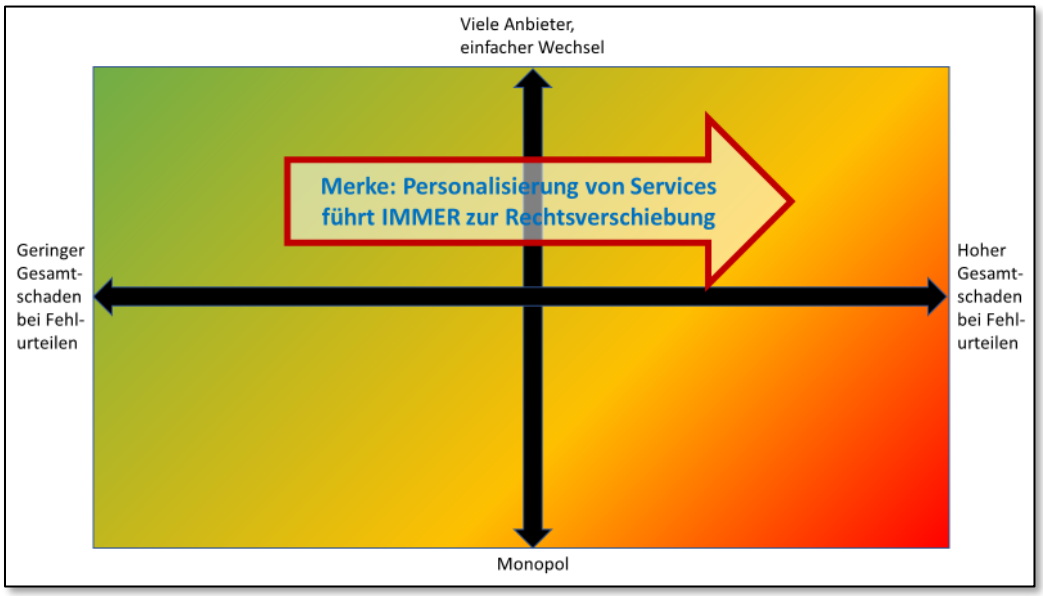


Wie kann nun ein ADM-System als Black-Box analysiert werden? Wir nehmen Googles Suchmaschine als Beispiel. Bis 2001 hat wohl niemand die Suchmaschine als besonders gesellschaftsgefährdend angesehen, daher ist sie zu Beginn als eher harmlos eingestuft worden. Das änderte sich mit Eli Parisers Buch "Die Filterblase". Darin beschreibt er, wie personalisierte Algorithmen dazu führen könnten, dass wir alle nur noch solche Art von Nachrichten präsentiert bekommen, die schon in unser politisches Bild passen. Er führt aus, welche Gefahren daraus erwachsen könnten. Die "Filterblase" wird zum stehenden Begriff auch in der Diskussion in Deutschland.



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

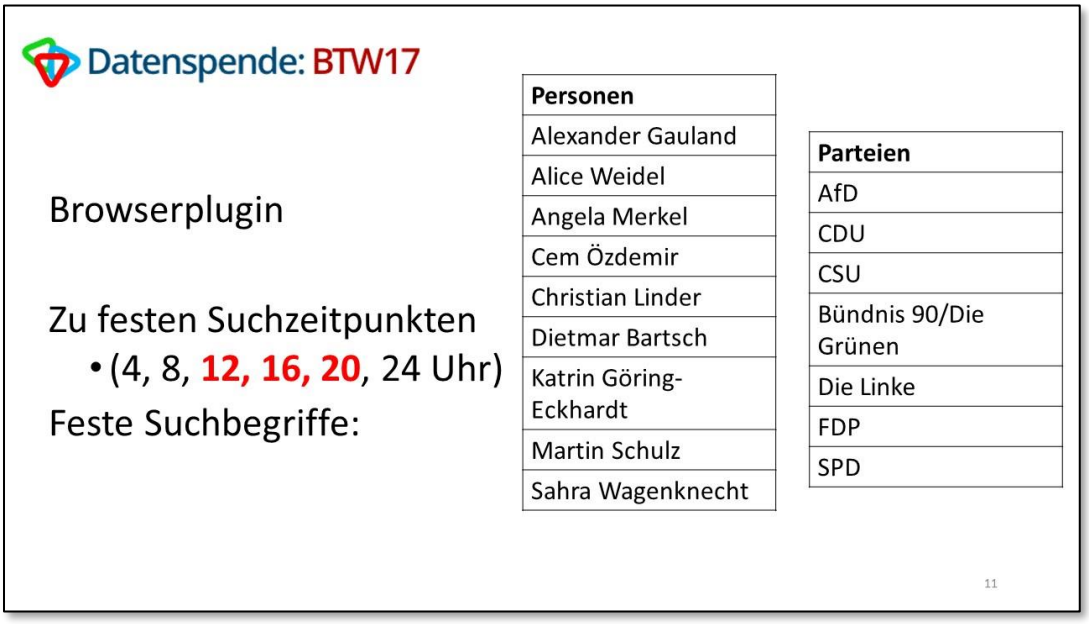
Parisers Buch basiert auf Anekdoten über das Personalisierungsverhalten von Algorithmen. Er zeigt in seinem TED-Talk diese zwei Suchergebnisse zum Thema "Ägypten" von zwei verschiedenen Personen. Diese sehen sehr unterschiedlich aus (einmal touristische Bilder, einmal eingeschobene Nachrichten), aber eigentlich sind von den gezeigten 4-5 "organischen" Suchergebnissen 3 gleich. Wie groß ist also der Personalisierungsgrad von Googles Suchergebnissen 2011? Das weiß keiner, es gibt keine empirischen Studien.



Grundsätzlich lässt sich aber festhalten, dass die Personalisierung eines algorithmischen Services **immer** zu einer Erhöhung des gesamtgesellschaftlichen Schadenspotenzials führt, da Gesellschaft dann viel schlechter überprüfen kann, wer welche Information bekommt. Dies ist relevant bei allen Produktempfehlungs-systemen, sozialen Netzwerken, Nachrichtenaggregatoren, ist ein Problem bei Kreditangeboten, Versicherungsangeboten, aber auch auf Plattformen wie stepstone, LinkedIn, Xing, etc.

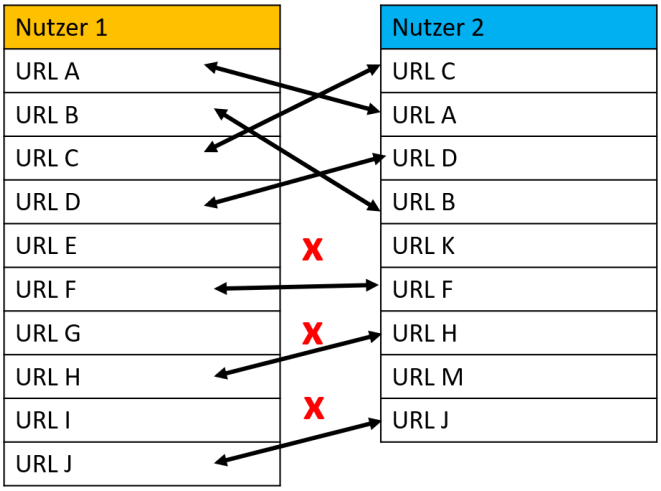
Bis 2017 wird viel diskutiert, aber niemand **misst** einfach mal, wie hoch denn der Personalisierungsgrad wirklich ist. Das ändert das Algorithm Accountability Lab zusammen mit AlgorithmWatch mit einem von den Landesmedienanstalten finanzierten Projekt[1]. Medienpartner war SpiegelOnline.

3 Beispiel für Black-Box-Analyse: Datenspende #BTW17



Interessierte Bürgerinnen und Bürger laden sich ein **Browser-Plugin** herunter, dass dann automatisch Suchanfragen zu festen Zeitpunkten macht zu festgelegten Schlüsselbegriffen (9 Politiker:innen, 7 Parteien). Die Bürgerinnen und Bürger "spenden" uns die Suchergebnisse der ersten Suchergebnisseite. Diese werden automatisch zentral gespeichert. 4384 mal wurde das Plugin heruntergeladen, wir haben insgesamt fast 6 Millionen gespendete Ergebnislisten.

[1] https://datenspende.algorithmwatch.org/



Für jedes Paar von Nutzern (zu demselben Zeitpunkt und zu demselben Suchbegriff) berechneten wir dann die Anzahl **nicht-geteilter** Links. Im Beispiel: Nutzer 1 hat 3 **nicht-geteilte Links** bezogen auf Nutzer 2 und Nutzer 2 hat nur 2 **nicht-geteilte Links** bezogen auf Nutzer 1.

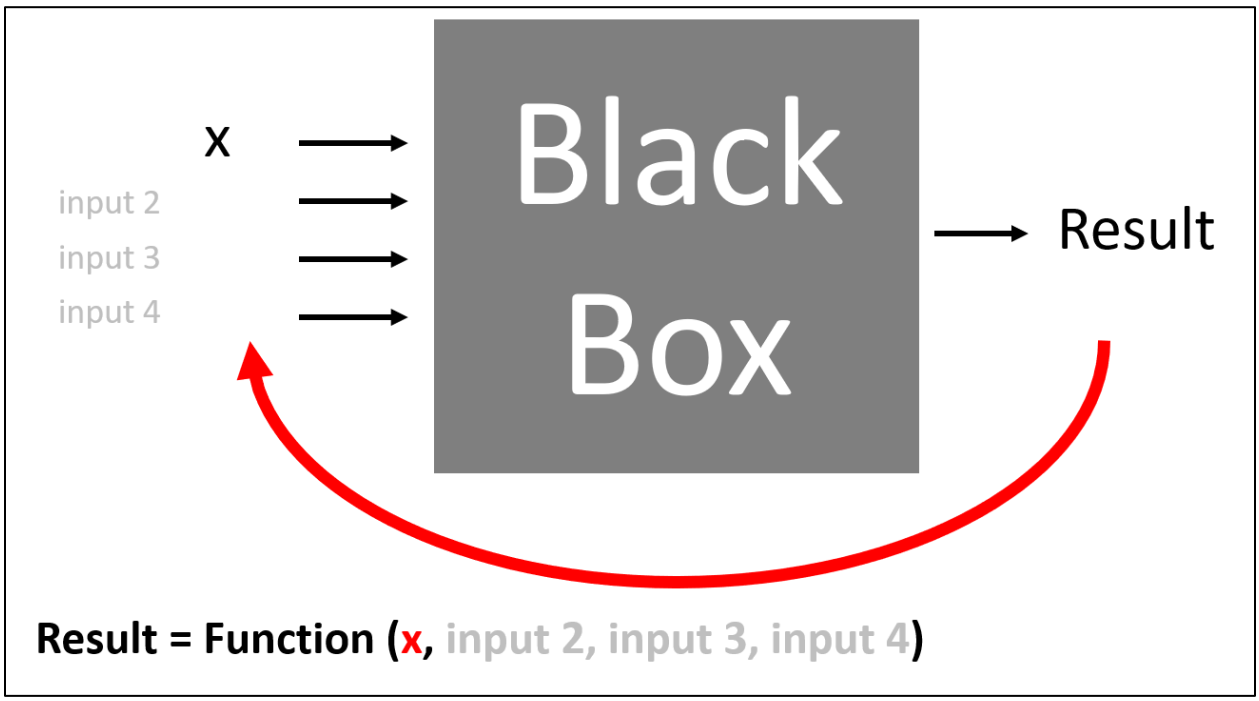
3.1 Ergebnis:

Bei Suchanfragen zu Politiker:innen waren im Durchschnitt nur 1-2 Links jeweils nicht geteilt. Bei Parteien im Durchschnitt 3-4 Links, davon waren aber viele regionaler Natur. D.h., ein Nutzer in Berlin bekommt für die SPD Links auf Berliner Ortsparteiwebseiten, eine Nutzerin in Kaiserslautern solche für Kaiserslautern und Umgebung. Zieht man diese ab, bleiben bei fast allen um die 2 Links im Durchschnitt nicht geteilt (außer bei AfD und CSU, da bleiben 2.9 und 2.7 Links übrig).

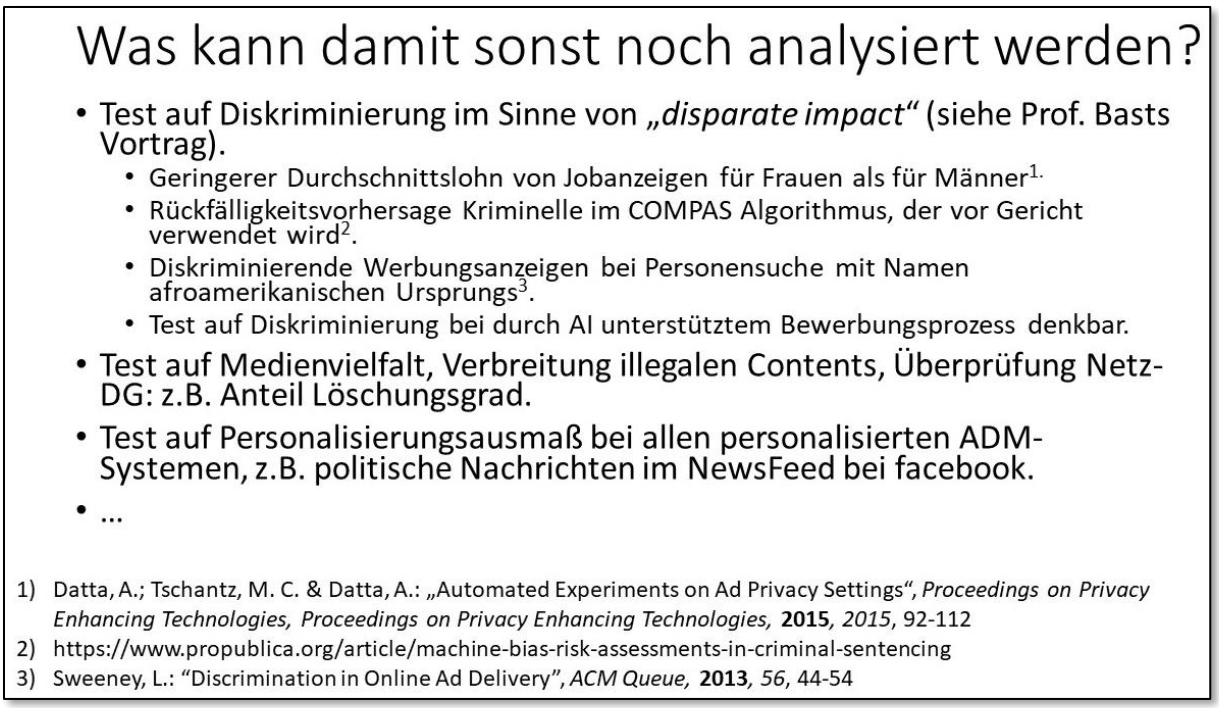
3.2 Kosten und Ort der Kontrolle:

Die Klassifizierung on Googles Suchmaschine in der Klasse 1 (in Bezug auf die Filterblasenthematik) unserer Risikomatrix würde auf der Ebene der technischen Regulierung eine "ständige Analyse" (des Personalisierungsgrades) erfordern. Was würde es kosten, wenn man eine solche Black-Box-Analyse ganzjährig mit wechselnden Suchanfragen durchführt? Die gesamte Studie hat ca. 120.000 € gekostet. Davon 45.000 € für die Software (wiederverwendbar!), 60.000 € für die Personenstunden. Da die Methode jetzt feststeht, wäre der Personalaufwand deutlich geringer! Wir schätzen den Aufwand pro Jahr auf 75.000 €. Eine der Landesmedienanstalten könnte dafür eine(n) Data Scientist einstellen – es bedarf dafür keiner neuen Kontroll-Institution.

4 Black-Box-Methode im Allgemeinen



Ganz allgemein wird das ADM-System bei der Black-Box-Methode als Berechnungssystem gesehen, dessen Mechanik unbekannt ist, dessen Verhalten (Resultat in Abhängigkeit von den Eingaben) jedoch beobachtet werden kann. Durch systematisches Variieren (möglichst jeweils nur einer) der Eingaben und den Einfluss dieser Eingabe auf das Resultat kann man versuchen, das Verhalten mathematisch-abstrakt zu beschreiben. Diese Methode stammt aus den empirischen Wissenschaften, ist aber auch (als "Testing") im Software-Engineering nicht unbekannt.



5 Was ist dazu notwendig?

- Unlimitierter und „anonymer“ Zugang zum AI-System
    - Der Anbieter darf nicht „wissen“, dass dies die Testanfragen sind
- Kenntnis über genaue Input-Struktur und Output-Struktur.
- Bei Personalisierungsgradanalysen, u.U. Datenspenden von ‚echten‘ Nutzern notwendig wegen Datenhistorie (**Typ „Datenspende“**).
    - Dann gezielter Zugriff auf die Informationen von Nutzer:innen notwendig, damit nicht Datensparsamkeit ermöglicht wird.
- Statt Daten von "echten" Nutzer:innen können für viele Fragestellungen „gefakte“ Nutzer simuliert werden.
    - Dann: Unlimitierte Anzahl zur Erstellung von nicht erkennbaren Fake-Accounts notwendig.

Diese Bedingungen sind z.B. bei Facebook beide nicht erfüllt: Nutzer:innen müssten uns ihren gesamten NewsFeed zur Verfügung stellen, auch wenn wir nur an politischer Werbung interessiert sind, da es keine Schnittstelle gibt, die uns

gezielten Zugriff erlaubt. Auch das unbegrenzte Anlegen von "Fake"-Nutzerkonten ist (aus guten Gründen) nicht möglich. Das führt dazu, dass der Personalisierungsgrad von Facebooks NewsFeed momentan nicht für eine relevante Anzahl von Nutzerkonten bestimmt werden könnte.

Dadurch steigt das Gesamtschadens**potenzial** von Facebooks Newsfeed, weil wir weniger untersuchen können und daher weiterhin vom Worst-Case ausgehen müssen. Es ist damit aus unserer Sicht mindestens ein Klasse 2-ADM-System, da es unter Umständen erheblichen Einfluss auf die (politische) Meinungsbildung hat.

6 Zusammenfassung

- Black-Box-Methode: Oft der einfachste Weg, um Verhalten von Algorithmen zu verstehen (es sei denn, Code ist sehr, sehr kurz).
- Transparenz von Code (=„Offenlegung“) nicht zielführend und oftmals schädlich (wirtschaftlich, aber auch wegen Manipulationsmöglichkeiten).
- Weitere Forschung zu Anwendungsbedingungen und Grenzen der Methodik notwendig.