

Explainable AI

Fairness und Transparenz in automatisierten Entscheidungen

13.12.2021

Dr. Timo Jakobi



Agenda

Das junge Feld der Verbraucherinformatik

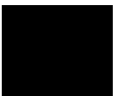
Exkurs Fairness und Explainable AI

Fallstudie Social Media

Erklärungsversuche und -bedarfe für automatisiertes Blocking auf Instagram

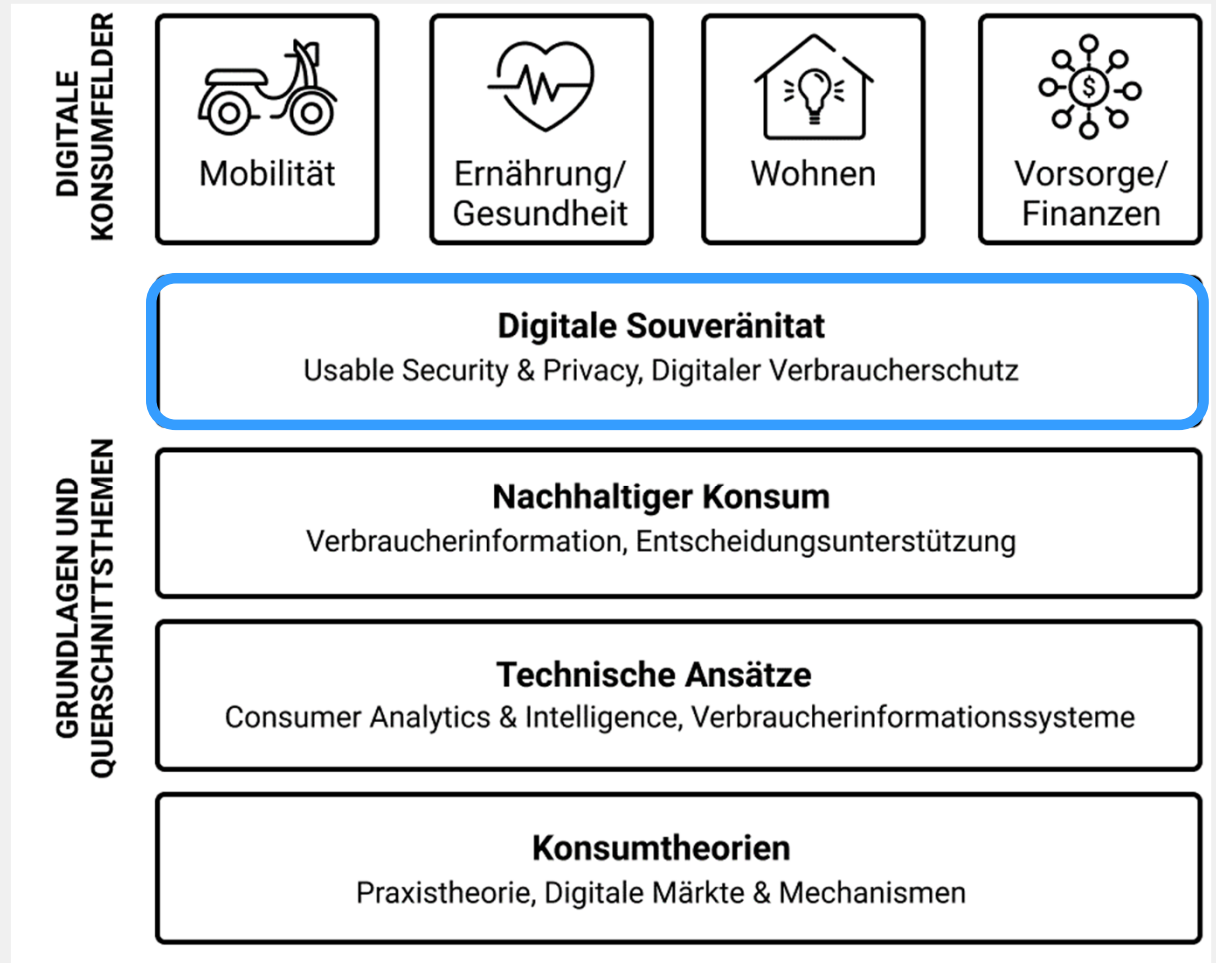
Fallstudie Versicherungstarife

(Un-)fares Design von Telematik-Tarifen aus Verbraucherperspektive

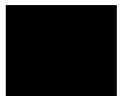
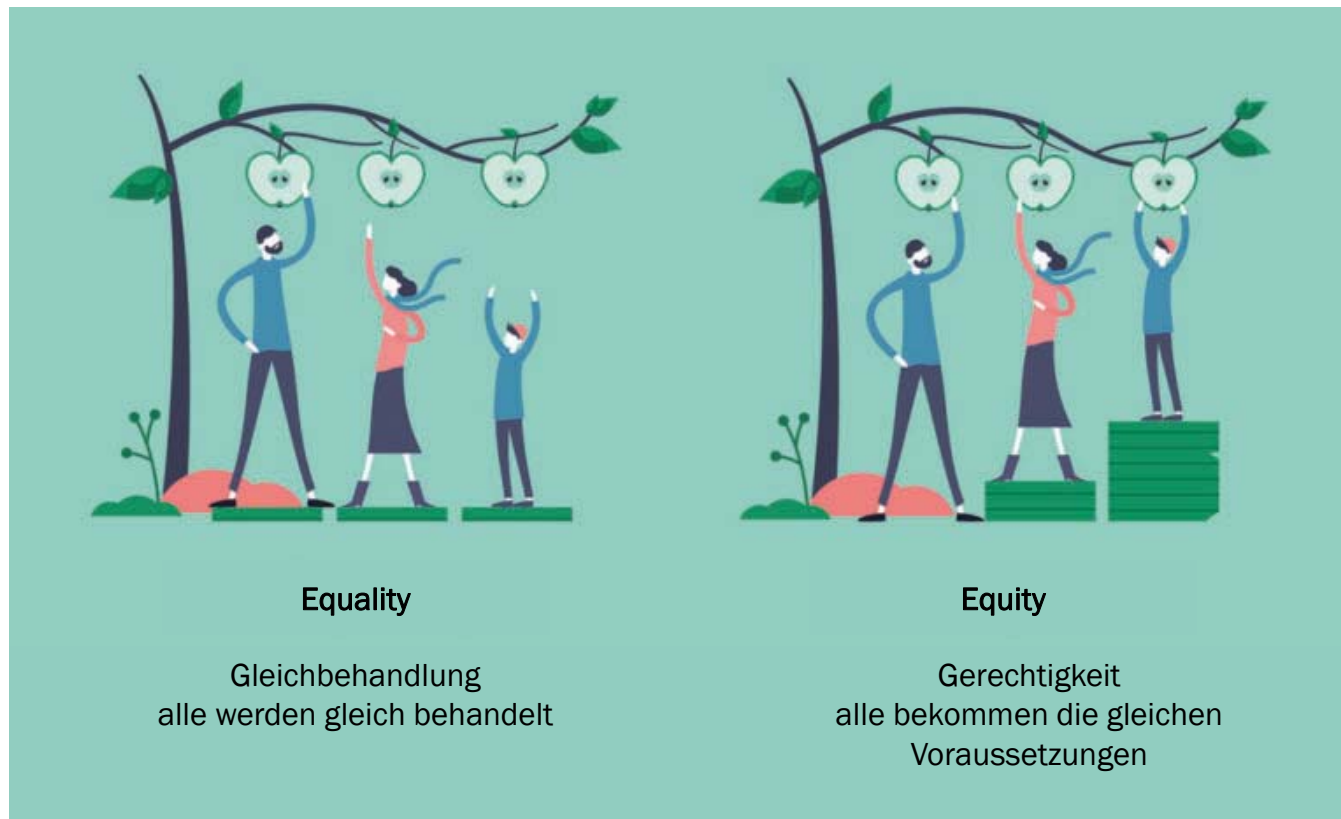


Forschungsfelder der Verbraucherinformatik

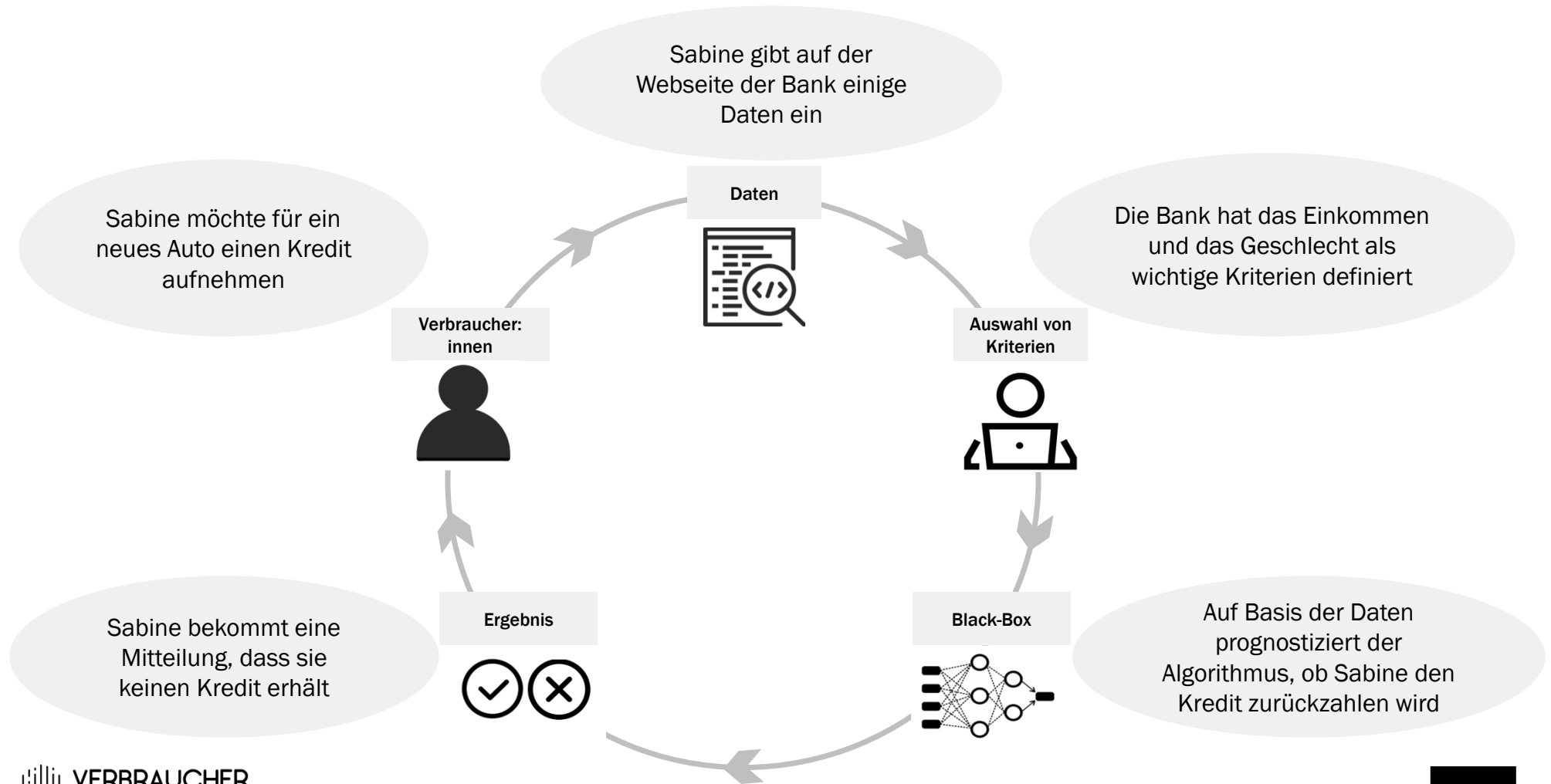
Untersuchung der zunehmenden Digitalisierung von Konsumfeldern und die Auswirkungen auf die Konsumpraktiken von Verbraucher*innen.



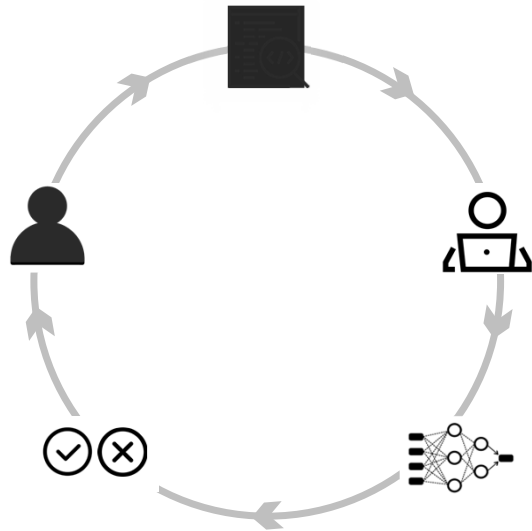
Was ist eigentlich Fairness?



Was sind eigentlich automatisierte Entscheidungen?



Fairness im Prozess automatisierter Entscheidungen



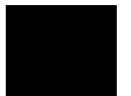
Welche Fairnessmaßstäbe haben Verbraucher: innen?

Wie finden wir „faire“ Daten?

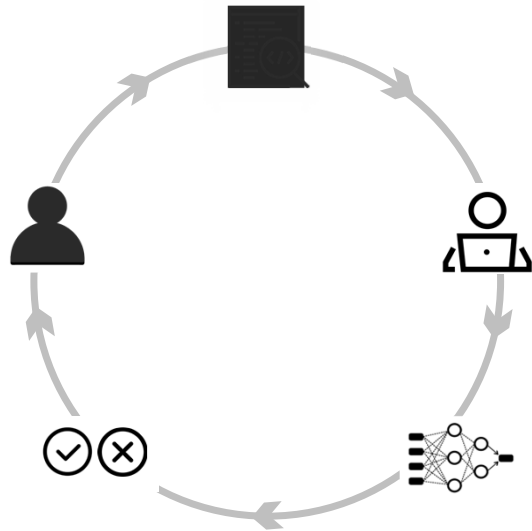
Welche Daten/Kriterien können wir einbeziehen?

Wie können wir den Prozess „fair“ gestalten?

Wann ist das Ergebnis „fair“?



Fairness im Prozess automatisierter Entscheidungen



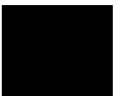
Welche Fairnessmaßstäbe haben Verbraucher: innen?

Wie finden wir „faire“ Daten?

Welche Daten/Kriterien können wir einbeziehen?

Wie können wir den Prozess „fair“ gestalten?

Wann ist das Ergebnis „fair“?





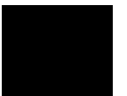
Was ist Explainable AI?

„Erklärbare“ KI (XAI) zielt darauf ab, Nutzer:innen in die Lage zu versetzen, das System und seine Funktionsweise zu verstehen



Fallstudie Social Media

Erklärungsversuche und -bedarfe für automatisiertes Blocking auf Instagram

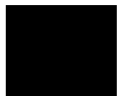


Social Media als Beispieltechnologie

- Weit verbreitet als
 - Informationsmedium
 - Kommunikationswerkzeug
 - Repräsentation des digitalen Ichs
 - Einkommensquelle
- Tendenziell hoher Einsatz von KI
 - Fake News
 - Sicherheitsmanagement

4,48

Mrd. Nutzer:innen weltweit





Methode

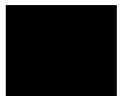
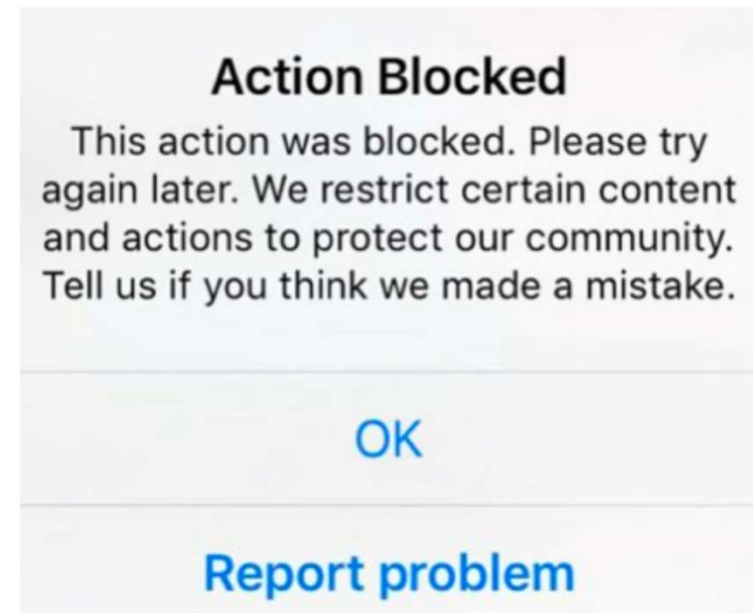
Fragestellung: Wie verstehen und erklären sich Nutzer:innen das automatische Blocking auf Instagram?

- 16 narrative Interviews
- Nutzer:innen von Instagram, die Blocking erfahren haben
- Deutscher Sprachraum, (10f, 6m), zwischen 19 und 57 Jahren alt



Ausgangslage: Unklare Blocking-Regeln

- Automatisiertes Blocking wird eingesetzt
- „Missbrauch“ ist generisch beschrieben
- Community klärt über mögliche Gründe für Blocking auf:
 - Typen
 - Action Ban,
 - Shadow Ban,
 - Account-Deaktivierung
 - Limits und Verhaltensweisen



Ergebnisse



Unklare Auslöser

”

[..] Ich wusste nicht, welche Aktivitäten zu einem Block führen könnten. Und ich musste mir das dann selber herleiten und raten, warum ich geblockt wurde, also vielleicht war das nicht so fair. Wenn man nicht informiert wird, ist das nicht fair.

”

[..] Also ich wusste, dass ich gebannt werden konnte, wenn ich schnell Leuten folgte. Aber ich dachte, wenn Instagram dir Profile vorschlägt, ist es in Ordnung, ihnen zu folgen, aber das stimmt nicht. Ich verstehe nicht, warum Instagram Menschen dazu ermutigt, etwas zu tun, was sie nicht tun dürfen.

- Konfligierende Ziele Sicherheit und Nutzer-Engagement
- Unklare Grenzen und Auslöser

Ergebnisse



Selbsterklärung und Lernfähigkeit nach Block

”

Ich habe alles überprüft. Ich habe den Beitrag überprüft, die Überschriften und ich bin zu anderen Stellen gegangen und habe andere Dinge ausprobiert

”

Wenn sie mir gesagt hätten, warum mir das passiert ist, hätte ich verhindern können, dass ich wieder blockiert werde, indem ich diese Aktion vermeide.

- Ex-Post Kommunikation
- Ursachenforschung wird dem Nutzer überlassen

Ergebnisse



Hilfestellung um zukünftige Blocks zu vermeiden

”

Ich habe es über diese Funktion gemeldet, aber es gab überhaupt keine Antwort. Ich weiß nicht einmal, ob diese Funktion was macht.

- Unklare Konfliktlösungs- und Interventionsmöglichkeiten

Ergebnisse



Gleichbehandlung

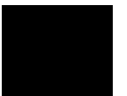
”

Das passiert normalen Menschen, oder kannst du dir vorstellen, dass sie Kendall Jenners Konto sperren?

- Inkonsistente Durchsetzung von Maßnahmen

Takeaway: Transparent kommunizieren

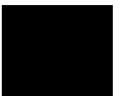
- Informationen auch VOR der Entscheidung über
 - Arten der automatisierten Entscheidungen
 - Mögliche Auswirkungen
 - Auslöser / Parameter
- Eigenschaften und Auswirkungen der Entscheidung erläutern
 - Dauer,
 - Umfang,
 - ...
- Kommunikationspfade für Einspruch





Fallstudie Versicherungstarife

(Un-)fares Design von Telematik-Tarifen aus Verbraucherperspektive



Anwendungsfall: Telematik Versicherungen

- Fahrverhalten wird ausgewertet
- Rabatt je nach persönlichem Fahrverhalten erhalten
- Typische Kriterien bspw. Bremsverhalten oder G-Kräfte in Kurven
- Werbeversprechen:
 - Umsichtiges Fahren wird belohnt,
 - „faire und transparente“ Bewertung



Eine gute Idee die Straßen etw... 25. Jan.



Mir :o)

...denn über den Geldbeutel lässt sich vieles erreichen. Eigentlich sollte dieses System flächendeckend eingeführt werden, dann wird der ein oder andere sicherlich erst überlegen bevor er aufs Gas steigt und andere Verkehrsteilnehmer gefährdet.

Einen Punkt Abzug gibt es dennoch, denn meiner Meinung nach ist das ganze noch nicht ganz ausgereift. Teilweise erfasst das System ganze Fahrten falsch (in der Stadt hat es die Parallelstraße erfasst) oder überhaupt nicht, zudem ist das Thema Handynutzung während der Fahrt so ein Thema. Meine Kleine schaut gerne Videos während der Fahrt (sie sitzt hinten) und das wird wohl negativ ausgewertet. Zudem erkennt das System eine Vollbremsung immer als schlecht, obwohl diese manchmal Leben rettet. Wenn man mit 70 km/h auf der Landstraße unterwegs ist und plötzlich ein Traktor auf die Straße zieht (dieser war hinter einer Kuppe versteckt) muss man stark bremsen.

Trotzdem an sich eine gute Idee.

Methode

Fragestellung: Wie nehmen Verbraucher: innen solche ihre Fahrscores in der Praxis wahr?

- 1003 Nutzerkommentare zur App „HUK - mein Auto“ aus dem Google Play Store analysiert
- Kommentare sind nicht nur ein guter Indikator für Benutzerfreundlichkeit und User Experience
- Viele der Kommentare beziehen sich auch auf Situationen, in denen Nutzer: innen aus Ihrer Perspektive „unfaire“ Bewertungen wiedergeben
- Inhaltsanalyse mit induktiver Kategorienbildung

Ergebnisse



Situationsbezogene Erklärungsansätze

”

*[..] Außerdem habe ich schon ein paar Mal erlebt, dass mein Bremsverhalten angezeigt wurde. In den Situationen musste ich stark bremsen, weil **Rehe kreuzten** oder weil ein **Kind** einen Ball von der Straße zurückholte. Normalerweise bin ich bei 85% in solchen Situationen wurde ich dann auf 79% runtergestuft.“*

- Nutzer:innen berichten von Erfahrungen aus ihrem Alltag und versuchen sich auf Basis dessen zu erklären.
- **Problematisch** sind Situationen, in denen die Reaktion des Systems für Verbraucher:innen nicht nachvollziehbar ist.

Ergebnisse



Fairnessbewertungen

”

[..] Aber ich denke, es wäre professioneller und fairer, wenn die Intelligenz in der App/Cloud solche Ausreißer erkennen und entsprechend aussortieren könnte.“

”

[..] Die App bekommt von mir einen Punkt Abzug, weil man in der App nicht gegen ungerechtfertigte negative Einträge einschreiten kann.“

- Moralisch richtiges Verhalten sollte nicht zur Abwertung im Score führen.
- Mangelnde Möglichkeit der Interaktion mit Anbieter und nicht zufriedenstellende Antworten.

Takeaways: Transparent kommunizieren

Vor der Nutzung



Bildung der Erwartungshaltung

Erwartungshaltung der Nutzer:
innen wird nicht erfüllt

Transparenz über den Zweck
des Systems

Während der Nutzung



Zielerreichung und Unterstützung

Verbraucher: innen verlieren die
„Kontrolle“ über das System

Alternativen Handlungsdiskurs

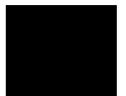
Nach der Nutzung



Korrigierbarkeit und Agency

Bewertung als „fehlerhaft“, obwohl
System das tut was es tun soll

Interaktions- und
Interventionsmöglichkeiten



Kontakt



Dr. Timo Jakobi

Institut für Verbraucherinformatik

Timo.Jakobi@h-brs.de

www.verbraucherinformatik.de

- Viele bestehende Dialogprinzipien sind in KI-Einsatz ebenso notwendig
 - Erwartungsmanagement
 - Lernfähigkeit der NutzerInnen unterstützen
 - Interventionsmöglichkeiten
- Welche Handlungen werden übernommen?
 - Was sind die Ziele/Zwecke?
 - Welche Daten werden berücksichtigt?
 - Ab wann greift eine KI ein?
 - Welche Maßnahmen dürfen ergriffen werden?
 - Was sind Limitierungen (Fairness?)

