sofia

Sonderforschungsgruppe
Institutionenanalyse

Die Tragik der Algorithm Aversion

Filiz, I.; Judek, J.R.; Lorenz, M.; Spiwoks, M.

sofia-Diskussionsbeiträge 21-2, Darmstadt 2021

ISBN: 978-3-941627-88-8

sofia-Diskussionsbeiträge
zur Institutionenanalyse
Nr.21-2

ISSN 1437-126X
ISBN 978-3-941627-88-8

Die Tragik der Algorithm Aversion

Ibrahim Filiz,

Jan René Judek,

Marco Lorenz und

Markus Spiwoks

Wolfsburg und Göttingen, Januar 2021

Keywords:
Algorithm aversion, technology adoption, framing, behavioral economics, experiments.

JEL classification:
D81, D91, G41, O33

Abstract:
Algorithmen bewältigen viele Aufgaben bereits zuverlässiger als menschliche Experten. Trotzdem zeigen einige Wirtschaftssubjekte eine ablehnende Haltung gegenüber Algorithmen (Algorithm Aversion). In manchen Entscheidungssituationen kann ein Fehler schwerwiegende Konsequenzen haben, in anderen Entscheidungssituationen nicht. Wir untersuchen im Rahmen eines Framing-Experimentes den Zusammenhang zwischen der Tragweite der Entscheidungssituation einerseits und der Häufigkeit der Algorithm Aversion andererseits. Dabei zeigt sich, dass die Algorithm Aversion umso häufiger auftritt, je gravierender die möglichen Konsequenzen einer Entscheidung sind. Gerade bei besonders wichtigen Entscheidungen führt somit die Algorithm Aversion zu einer Reduzierung der Erfolgswahrscheinlichkeit. Das kann man als die Tragik der Algorithm Aversion bezeichnen.

Ibrahim Filiz
Ostfalia University of Applied Sciences, Faculty of Business, Siegfried-Ehlers-Str. 1, D-38440 Wolfsburg, Germany, Tel.: +49 160 3344 078
E-Mail: ibrahim.filiz@ostfalia.de

Jan René Judek
Ostfalia University of Applied Sciences, Faculty of Business, Siegfried-Ehlers-Str. 1, D-38440 Wolfsburg, Germany, Tel.: +49 5361 892 225 420
E-Mail: ja.judek@ostfalia.de

Marco Lorenz
Georg August University Göttingen, Faculty of Economic Sciences, Platz der Göttinger Sieben 3, D-37073 Göttingen, Germany, Tel.: +49 1522 6672 503
E-Mail: marco.lorenz@stud.uni-goettingen.de

Markus Spiwoks
Ostfalia University of Applied Sciences, Faculty of Business, Siegfried-Ehlers-Str. 1, D-38440 Wolfsburg, Germany, Tel.: +49 5361 892 225 100
E-Mail: m.spiwoks@ostfalia.de

Inhalt

1 Einführung 5
2 Experimentelles Design und Hypothesen 11
3 Ergebnisse 14
4 Zusammenfassung 20
5 Literatur 22
6 Anhang 26
6.1 Anhang 1: Spielanleitung 26
6.2 Anhang 2: Kontrollfragen 27
6.3 Anhang 3: Entscheidungssituationen Treatment A und B 28
6.4 Anhang 4: Ermittlung des Zufallsereignisses mit Hilfe einer Lotterie 30
6.5 Anhang 5: Definitionen Algorithm Aversion in der Literatur 31

Abbildungen

Abbildung 1: Anteile der Probanden hinsichtlich der Zugehörigkeit zu einer Fakultät bzw. des Geschlechts bei Aufteilung auf die Treatments A und B 15
Abbildung 2: Boxplot zur Einschätzung der Tragweite der Entscheidungssituationen 16
Abbildung 3: Entscheidungen zugunsten des Algorithmus in Abhängigkeit von der Tragweite der Entscheidungssituation 18
Abbildung 4: Kartenstapel bei Wahl des Algorithmus 30
Abbildung 5: Kartenstapel bei Wahl des menschlichen Experten 30

Tabellen

Tabelle 1: Definitionen der Algorithm Aversion in der Literatur 6
Tabelle 2: Treatments und Entscheidungssituationen 11
Tabelle 3: Einschätzung der Tragweite in den Treatments A und B 15
Tabelle 4: Entscheidungen für und gegen den Algorithmus 17

1
Einführung

Automatisierte Entscheidungswege oder Entscheidungshilfen, so genannte Algorithmen, gewinnen zusehends an Bedeutung für den beruflichen und den privaten Alltag vieler Menschen. Insbesondere die fortschreitende Digitalisierung und die wachsende Bedeutung der künstlichen Intelligenz haben dazu geführt, dass uns bereits seit einigen Jahrzehnten leistungsfähige Algorithmen zur Verfügung stehen (vgl. bspw. Dawes, Faust & Meehl, 1989). Diese Algorithmen bewältigen viele Aufgaben bereits zuverlässiger als menschliche Experten. Aber nur wenige Algorithmen arbeiten vollkommen fehlerfrei. Manche Anwendungsbereiche von Algorithmen haben im Falle eines Fehlers schwerwiegende Folgen – wie beispielsweise das autonome Fahren (vgl. Shariff, Bonnefon & Rahwan, 2017), die Erstellung medizinischer Diagnosen (vgl. Majumdar & Ward, 2011) oder die Unterstützung bei Strafverfahren (vgl. Simpson, 2016). Andererseits werden Algorithmen auch für Tätigkeiten eingesetzt, die im Falle eines Fehlers keine derart gravierenden Folgen haben – wie beispielsweise die Partnervermittlung (vgl. Brozovsky & Petříček, 2007), die Prognose des Wetters (vgl. Sawaitul, Wagh & Chatur, 2012) und die Empfehlung von Kochrezepten (vgl. Ueda, Takahata & Nakajima, 2011).

Einige Wirtschaftssubjekte zeigen eine ablehnende Haltung gegenüber Algorithmen, die üblicherweise als Algorithm Aversion bezeichnet wird (für eine Übersicht zur Algorithm Aversion siehe Burton, Stein & Jensen, 2020). Viele Entscheidungsträger neigen deshalb dazu, Aufgaben eher an menschliche Experten zu delegieren oder sie selbst auszuführen. Das trifft oft selbst dann zu, wenn klar erkennbar ist, dass der Einsatz von Algorithmen zu einer Steigerung der Ergebnisqualität beitragen würde.

Bisherigen Publikationen zu diesem Themenbereich definieren den Begriff der Algorithm Aversion durchaus unterschiedlich (Tabelle 1). Diese unterschiedlichen Begriffsverständnisse spiegeln sich sowohl in den Argumentationslinien als auch in den experimentellen Designs wider. Aus Sicht einiger Forscher kann erst dann von Algorithm Aversion gesprochen werden, wenn der Algorithmus erkennbar die Entscheidungsalternative mit der höchsten Ergebnisqualität oder Erfolgswahrscheinlichkeit darstellt (vgl. Burton, Stein & Jensen, 2020; Köbis & Mossink, 2020; Castelo, Bos & Lehmann, 2019; Ku, 2020; Dietvorst, Simmons & Massey, 2018; Dietvorst, Simmons & Massey, 2015).

Andere Autoren sprechen hingegen bereits von Algorithm Aversion, sobald Wirtschaftssubjekte ungeachtet einer etwaigen Überlegenheit des Algorithmus eine grundsätzliche Ablehnung gegenüber diesem offenbaren (vgl. Efendić, Van de Calseyde & Evans, 2020; Niszczota & Kaszás, 2020; Horne et al., 2019; Logg, Minson & Moore, 2019; Rühr et al., 2019; Yeomans et al., 2019; Prahl & Van Swol, 2017).

Ein weiterer wichtiger Aspekt beim Begriffsverständnis der Algorithm Aversion ist die Frage, ob und gegebenenfalls wie die Wirtschaftssubjekte von der Überlegenheit des Algorithmus erfahren. In bisherigen Studien werden unterschiedliche Ansätze gewählt. Dietvorst, Simmons und Massey (2015) etablieren das Sammeln von Erfahrung im Umgang mit einem Algorithmus, um dessen Erfolgswahrscheinlichkeit im Vergleich zur eigenen Leistung einschätzen zu können. Ebenfalls Dietvorst, Simmons und Massey (2018) geben den durchschnittlichen Fehler des Algorithmus an. Alexander, Blinder und Zak (2018) machen exakte Angaben über die Erfolgswahrscheinlichkeit des Algorithmus oder sie weisen auf die Nutzungsrate des Algorithmus durch andere Wirtschaftssubjekte in der Vergangenheit hin.

Tabelle 1: Definitionen der Algorithm Aversion in der Literatur

Autoren	**Definition der Algorithm Aversion (deutsche Übersetzung)**
Dietvorst, Simmons & Massey, 2015	"Untersuchungen zeigen, dass evidenzbasierte Algorithmen die Zukunft genauer vorhersagen als menschliche Prognostiker. Wenn Prognostiker jedoch entscheiden, ob sie einen menschlichen Prognostiker oder einen statistischen Algorithmus verwenden, wählen sie häufig den menschlichen Prognostiker. Dieses Phänomen, das wir *Algorithm Aversion* nennen, (…)"
Prahl & Van Swol, 2017	"Die irrationale Abwertung von systemseitigen Empfehlungen ist seit langem bekannt und eine Quelle der lebhaften Debatte bezüglich subjektiv-menschgemachter und datenbasierter Entscheidungsmodelle in der psychologischen Forschung (Dawes, 1979; Meehl, 1954). Kürzlich wurde dieser Effekt in der Prognoseforschung festgestellt (Önkal et al., 2009) und wurde als Algorithm Aversion bezeichnet (Dietvorst, Simmons & Massey, 2015)."
Dietvorst, Simmons & Massey, 2018	"Obwohl evidenzbasierte Algorithmen menschliche Prognostiker regelmäßig übertreffen, werden sie von Entscheidungsträgern häufig nicht verwendet, sobald diese feststellen, dass ein Algorithmus nicht perfekt ist - ein Phänomen, das als *Algorithm Aversion* bekannt ist."
Castelo, Bos & Lehmann, 2019	"Der Siegeszug der Algorithmen hat zur Folge, dass Nutzer zunehmend vor einer neuartigen Entscheidung stehen: Sollten sie sich mehr auf Menschen oder auf Algorithmen verlassen? Untersuchungen legen nahe, dass die Standardoption bei dieser Wahl darin besteht, sich auf den Menschen zu verlassen, selbst wenn dies objektiv betrachtet zu schlechteren Ergebnissen führt."
Commerford, Dennis, Joe & Wang, 2019	"(…) *Algorithm Aversion* - die Tendenz von Einzelpersonen, computergestützte Ratschläge weniger stark zu berücksichtigen als menschliche Ratschläge, obwohl die Ratschläge ansonsten identisch sind."
Horne, Nevo, O'Donovan, Cho & Adali, 2019	„Zum Beispiel untersuchen Dietvorst et al. (Dietvorst, Simmons und Massey 2015), wann Menschen den menschlichen Prognostiker einem statistischen Algorithmus vorziehen. Die Autoren stellen fest, dass die Abneigung gegen das automatisierte Tool zunimmt, wenn ihm die Menschen bei der Arbeit zusehen können, selbst wenn der Algorithmus signifikant bessere Ergebnisse hervorbringt als der Mensch. Dietvorst et al. erklären, dass die Abneigung aufgrund einer schnelleren Abnahme des Vertrauens in algorithmische als in menschliche Prognostiker entsteht, wenn beide denselben Fehler begehen (Dietvorst, Simmons und Massey 2015). "
Ku 2019	""Algorithm Aversion", ein Begriff, auf den sich Dietvorst et al. (Dietvorst et al. 2015) beziehen, bedeutet, dass Menschen dem Algorithmus misstrauen, obwohl der Algorithmus den Menschen durchweg übertrifft."

Leyer & Schneider, 2019	"Im spezifischen Kontext der Delegation von Entscheidungen an KI-fähige Systeme haben jüngste Erkenntnisse eine allgemeine Algorithm Aversion aufgedeckt, eine irrationale Vernachlässigung solcher Systeme als geeignete Entscheidungsträger entgegen objektiver Anhaltspunkte (Dietvorst, Simmons und Massey, 2018)."
Logg, Minson & Moore, 2019	"(...) menschliches Misstrauen gegenüber dem Output eines Algorithmus, manchmal auch als "Algorithm Aversion" bezeichnet (Dietvorst, Simmons & Massey, 2015).[1] "; Fußnote 1: "Obwohl es in diesem einflussreichen Artikel [von Dietvorst et al.] um die Auswirkungen des Beobachtens eines algorithmischen Prognosefehlers auf die Wahrscheinlichkeit, den Algorithmus erneut auszuwählen, geht, wird er teilweise zitiert, als ginge es darum, wie häufig Menschen ganz allgemein auf einen Algorithmus zurückgreifen."
Önkal, Gönül & De Baets, 2019	„(…) Menschen sind abgeneigt, Ratschläge von Algorithmen zu verwenden, und verzeihen keine Fehler, die durch den Algorithmus gemacht werden (Dietvorst et al., 2015; Prahl & Van Swol, 2017). "
Rühr, Streich, Berger & Hess, 2019	"Es wurde gezeigt, dass Benutzer eine Abneigung gegen algorithmische Entscheidungssysteme [Dietvorst, Simmons, Massey, 2015] sowie gegen den wahrgenommenen Kontrollverlust zeigen, der mit einer übermäßigen Delegation von Entscheidungshoheit verbunden ist [Dietvorst, Simmons, Massey, 2018]."
Yeomans, Shah, Mullainathan & Kleinberg, 2019	"(...) Menschen bevorzugen es, Empfehlungen von einem Menschen zu erhalten, anstatt von einem Empfehlungssystem (...). Dies spiegelt jahrzehntelange Forschung wider, die zeigt, dass Menschen es ablehnen, sich auf Algorithmen zu verlassen, wobei der Hauptgrund der Abneigung algorithmische Fehler sind (für eine Übersicht siehe Dietvorst, Simmons & Massey, 2015)"
Berger, Adam, Rühr & Benlian, 2020	"Frühere Untersuchungen haben jedoch gezeigt, dass Menschen die Unterstützung durch Menschen häufig der Unterstützung durch ein IT-System vorziehen, selbst wenn letzteres eine überlegene Leistung bietet - ein Phänomen, das als Algorithm Aversion bezeichnet wird." (…)„Diese Unterschiede führen zu zwei verschiedenen Ansichten darüber, was unter Algorithm Aversion zu verstehen ist: Unwilligkeit, sich auf einen Algorithmus zu verlassen, den ein Benutzer als fehlerhaft empfunden hat, oder allgemeiner Widerstand gegen algorithmische Einschätzungen."
Burton, Stein & Jensen, 2020	"(…) Algorithm Aversion - die Ablehnung menschlicher Prognostiker, überlegene, aber nicht perfekte Algorithmen zu verwenden - (…)"
De-Arteaga, Fogliato & Chouldechova, 2020	„*Algorithm Aversion* - die Tendenz, Empfehlungen einer Maschine zu ignorieren nachdem man gesehen hat, dass diese fehlerhaft sein können (…)"
Efendić, Van de Calseyde & Evans, 2020	"Algorithmen funktionieren bei verschiedenen Vorhersageaufgaben durchweg gut, aber Menschen misstrauen oft ihren Ratschlägen."; „Wiederholte Beobachtungen zeigen jedoch, dass die Menschen Ratschlägen eines Algorithmus zutiefst misstrauen, insbesondere wenn sie dem Algorithmus bei einem Fehler zusehen können. (Bigman & Gray, 2018; Diab, Pui, Yankelevich & Highhouse, 2011; Dietvorst, Simmons & Massey, 2015; Önkal, Goodwin, Thomson, Gönül & Pollock, 2009). "
Erlei, Nekdem, Meub, Anand & Gadiraju, 2020	"In letzter Zeit hat das Thema Algorithm Aversion großes Interesse geweckt (siehe (Burton, Stein und Jensen 2020) für eine Übersicht). In ihrer wegweisenden Studie zeigen (Dietvorst, Simmons und Massey 2015), dass menschliche Akteure aus der Beobachtung algorithmischer Fehler anders lernen als aus der Beobachtung menschlicher Fehler. Sogar Teilnehmer, die direkt mit ansehen konnten, wie ein Algorithmus bessere Ergebnisse geliefert hat als ein Mensch, verwendeten das Modell seltener, nachdem sie seine Unvollkommenheiten beobachtet hatten."

Germann & Merkle, 2020	"Die Tendenz des Menschen, die Verwendung von Algorithmen zu scheuen, selbst wenn Algorithmen ihr menschliches Gegenstück deutlich übertreffen, wurde als Algorithm Aversion bezeichnet."
Ireland, 2020	„(…) einige Forscher stellen fest, dass Personen gegenüber Algorithmen im Vergleich zu Menschen abgeneigt sind, nachdem sie Fehler gleichen Ausmaßes beobachtet haben.“
Jussupow, Benbasat & Heinzl, 2020	"(…) die Literatur legt nahe, dass Algorithmen zwar häufig eine überlegene Leistung aufweisen, Benutzer jedoch nur ungern mit Algorithmen anstelle von menschlichen Agenten interagieren - ein Phänomen, das als Algorithm Aversion bekannt ist."
Kawaguchi, 2020	„Das Phänomen, bei dem Menschen häufig unterlegenen menschlichen Entscheidungen folgen, sogar wenn sie verstehen, dass algorithmische Entscheidungen diese übertreffen, ist weit verbreitet. Es wird als Algorithm Aversion bezeichnet (Dietvorst et al. 2015). “
Köbis & Mossink, 2020	„Wenn Menschen von einem Algorithmus erfahren, zeigen umfangreiche Untersuchungen, dass sie gegenüber algorithmischen Entscheidungsträgern im Allgemeinen abgeneigt sind. Diese Ablehnung von „menschlichen Entscheidungsträgern, überlegene, aber unvollkommene Algorithmen zu verwenden“ (Burton, Stein & Jensen, 2019; S. 1) wurde als Algorithm Aversion bezeichnet (Dietvorst, Simmons & Massey, 2015). Zum Teil aufgrund der Annahme, dass menschliche Fehler zufällig verteilt sind, während algorithmische Fehler systematisch sind (Highhouse, 2008), haben die Menschen in verschiedenen Bereichen eine Ablehnungshaltung gegen Algorithmen gezeigt (für eine systematische Literaturübersicht siehe Burton et al., 2019). “
Niszczota & Kaszás, 2020	„Wenn Menschen die Möglichkeit haben, zwischen Ratschlägen eines Menschen oder eines Algorithmus zu wählen, bevorzugen sie erstere Option und zeigen somit eine Algorithm Aversion. (Castelo et al., 2019; Dietvorst et al., 2015, 2016; Longoni et al., 2019). ”
Wang, Harper & Zhu, 2020	"(...) Menschen neigen dazu, Menschen mehr zu vertrauen als Algorithmen, selbst wenn der Algorithmus genauere Vorhersagen macht."

Die englischsprachigen Originalzitate befinden sich in Anhang 5.

Zudem ist die Ausgestaltung des Feedbacks im Umgang mit Algorithmen von Bedeutung. Können Wirtschaftssubjekte (unter Einbeziehung ihrer bisherigen Entscheidungen) einen Rückschluss auf die Qualität bzw. den Erfolg eines Algorithmus ziehen? Dietvorst, Simmons und Massey (2015) verwenden Feedback lediglich, um Erfahrung im Umgang mit einem Algorithmus zu ermöglichen. Prahl und Van Swol (2017) geben Feedback nach jeder einzelnen Entscheidung, wodurch die Einschätzung der Erfolgsträchtigkeit des Algorithmus möglich wird. Filiz et al. (2021) folgen diesem Ansatz und verwenden Feedback nach jeder einzelnen Entscheidung, um den Rückgang der Algorithm Aversion im Zeitverlauf zu untersuchen.

Weitere Gesichtspunkte, die aus den bisherigen Definitionen der Algorithm Aversion in der Literatur hervorgehen, sind die Zuverlässigkeit des Algorithmus (perfekt oder unvollkommen), die Beobachtung der Zuverlässigkeit (sichtbares Auftreten von Fehlern), der Zugriff auf historische Daten zur Erstellung der algorithmischen Prognose, das Setting (Algorithmus vs. Experte; Algorithmus vs.

Laie; Algorithmus vs. handelndes Wirtschaftssubjekt) sowie das Ausmaß des algorithmischen Eingriffs (Algorithmus fungiert als Entscheidungshilfe oder führt die Tätigkeit selbstständig aus).

Unserer Auffassung nach sind die Überlegenheit des Algorithmus (höhere Erfolgswahrscheinlichkeit) und die Kenntnis dieser Überlegenheit die entscheidenden Aspekte. Wir sprechen nur dann von Algorithm Aversion, wenn ein Wirtschaftssubjekt genau weiß, dass ein Verzicht auf den Algorithmus den Erwartungswert seines Nutzens reduziert, und trotzdem auf den Einsatz des Algorithmus verzichtet. Eine Entscheidung gegen einen bekanntermaßen überlegenen Algorithmus verringert den Erwartungswert des pekuniären Nutzens und muss folglich als Verhaltensanomalie (vgl. Frey, 1992; Kahneman & Tversky, 1979; Tversky & Kahneman, 1974) betrachtet werden.

In Entscheidungssituationen, die im Falle eines Fehlers weniger schwerwiegende Folgen nach sich ziehen, hat eine solche Verhaltensanomalie nicht allzu große Auswirkungen. In der Partnervermittlung kommt es schlimmstenfalls zu einem unverbindlichen Treffen mit einem unpassenden Kandidaten. Bei einer fehlerhaften Wetterprognose wird schlimmstenfalls unpassende Kleidung gewählt und bei der Empfehlung von Kochrezepten schmeckt schlimmstenfalls das Essen nicht. Doch insbesondere bei Entscheidungen, die im Falle eines Fehlers schwerwiegende Konsequenzen nach sich ziehen, wäre eine vom Rationalkalkül abweichende Entscheidung sehr riskant. So zieht ein Autounfall oder eine medizinische Fehldiagnose im Extremfall den Tod eines Menschen nach sich. Eine Verurteilung in einem Strafverfahren kann eine langjährige Haftstrafe zur Folge haben. In diesen schwerwiegenden Fällen sollte unbedingt Vernunft walten und ein im Hinblick auf die Erfolgswahrscheinlichkeit überlegener Algorithmus Verwendung finden. Kann eine Algorithm Aversion in gravierenden Situationen überwunden werden, um eine nutzenmaximierende Entscheidung zu treffen und damit bestenfalls Leben zu retten?

Tversky & Kahneman (1981) zeigen, dass Entscheidungen maßgeblich vom kontextualen Rahmen der Entscheidungssituation beeinflusst werden können. Die zur Veranschaulichung der Problemstellung gewählte Geschichte beeinflusst die Salienz der Informationen, was auch zu einer irrationalen Vernachlässigung der zugrundeliegenden mathematischen Gegebenheiten führen kann. Dieses Phänomen wird auch als Framing-Effekt bezeichnet (für eine Übersicht siehe Cornelissen & Werner, 2014). Wirtschaftssubjekte lassen sich, ungeachtet der tatsächlichen Erfolgswahrscheinlichkeiten, in ihrem Verhalten beeinflussen. Eine Algorithm Aversion könnte in unterschiedlichen Entscheidungskontexten verschieden stark ausgeprägt sein. Es ist denkbar, dass Wirtschaftssubjekte, die über den Einsatz eines Algorithmus entscheiden müssen, auch die Tragweite der Entscheidung berücksichtigen. In dieser Studie wird deshalb unter Verwendung eines Framing-Ansatzes untersucht, ob Wirtschaftssubjekte bereit sind, in Entscheidungssituationen, die schwerwiegende Konsequenzen nach sich ziehen können, von ihrer Algorithm Aversion abzulassen. Es wird also betrachtet,

ob es zu signifikant unterschiedlichen Häufigkeiten der Algorithm Aversion kommt – je nachdem, ob Entscheidungssituationen schwerwiegende Konsequenzen haben können oder nicht.

2
Experimentelles Design und Hypothesen

Zur Beantwortung der Forschungsfrage führen wir ein ökonomisches Experiment durch. Dabei nehmen die Probanden die Perspektive eines Unternehmers ein, der eine Dienstleistung für seine Kunden anbietet. Es steht die Entscheidung an, ob diese Dienstleistung von spezialisierten Algorithmen oder von menschlichen Experten ausgeführt werden soll.

In diesem Framing-Ansatz werden drei Entscheidungssituationen mit möglicherweise gravierenden Konsequenzen (Treatment A) und drei Entscheidungssituationen mit deutlich weniger gravierenden Auswirkungen (Treatment B) gegenübergestellt. Im Treatment A handelt es sich um folgende Dienstleistungen: (1) Fahrdienstleistungen mit Hilfe von autonomen Fahrzeugen (Algorithmus) oder mit Hilfe von Chauffeuren, (2) Auswertung von MRT-Scans mit Hilfe eines spezialisierten Computerprogramms (Algorithmus) oder mit Hilfe von Medizinern und (3) Auswertungen von Strafverfahrensakten mit Hilfe eines spezialisierten Computerprogramms (Algorithmus) oder mit Hilfe von Juristen. Im Treatment B handelt es sich um folgende Dienstleistungen: (1) Partnervermittlung auf einer Dating-Plattform mit Hilfe eines spezialisierten Computerprogramms (Algorithmus) oder mit Hilfe von psychologisch geschulten Mitarbeitern, (2) Auswahl von Kochrezepten für zu liefernde Kochboxen mit Hilfe eines spezialisierten Computerprogramms (Algorithmus) oder mit Hilfe von kulinarisch geschulten Mitarbeitern und (3) Erstellung von Wetterprognosen mit Hilfe eines spezialisierten Computerprogramms (Algorithmus) oder mit Hilfe von erfahrenen Meteorologen (Tabelle 2).

Tabelle 2: Treatments und Entscheidungssituationen

Entscheidungssituation	**Treatment**
Autonomes Fahren Auswertung von MRT-Scans Aktenauswertung bei Strafverfahren	A (möglicherweise gravierende Auswirkungen)
Partnervermittlung Auswahl von Kochrezepten Erstellung von Wetterprognosen	B (keine gravierenden Auswirkungen)

Die Entscheidungssituationen sind so ausgewählt, dass sie den Probanden aus der öffentlichen Debatte oder aus eigener Erfahrung bekannt sein sollten. Es wird den Probanden somit erleichtert, sich in den jeweiligen Kontext hineinzuversetzen. Die ausformulierten Entscheidungssituationen befinden sich im Anhang 3.

Die Studie ist als Between-Subjects-Design konzipiert. Jeder Proband wird mit nur einer der insgesamt sechs Entscheidungssituationen konfrontiert. Bei allen sechs Entscheidungssituationen liegen dieselben Erfolgswahrscheinlichkeiten vor: Der Algorithmus führt die Dienstleistung mit einer Wahrscheinlichkeit von 70% erfolgreich durch. Der menschliche Experte führt die Dienstleistung mit einer Wahrscheinlichkeit von 60% erfolgreich durch. Darüber hinaus ist auch die Vergütungsstruktur in beiden Treatments identisch. Die Teilnehmer erhalten eine Antrittsprämie (show-up fee) in Höhe von 2,00 €. Ferner wird eine zusätzliche Vergütung in Höhe von 4,00 € gezahlt, sofern die Dienstleistung erfolgreich durchgeführt wird. Lediglich der kontextuale Rahmen der sechs Entscheidungssituationen variiert.

Zunächst sind die Probanden aufgefordert, die Tragweite der Entscheidungssituation auf einer Skala von 0 (nicht gravierend) bis 10 (sehr gravierend) einzuschätzen. Diese Angabe dient als Manipulation Check. So wird erkennbar, ob die Probanden die Tragweite der Entscheidungssituationen im Treatment A tatsächlich als gravierender wahrnehmen als die Tragweite der Entscheidungssituationen im Treatment B. Beim autonomen Fahren und bei der Auswertung von MRT-Scans geht es unter Umständen um Kopf und Kragen. Bei der Auswertung von Unterlagen im Rahmen von Strafverfahren geht es unter Umständen um schwerwiegende Einschränkungen der persönlichen Freiheit. Die drei Entscheidungssituationen des Treatments A können bei einem ungünstigen Ausgang also gravierende Konsequenzen für Dritte haben. Anders sieht es bei der Partnervermittlung, bei der Auswahl von Kochrezepten und bei der Erstellung von Wetterprognosen aus. Selbst wenn diese Aufgaben gelegentlich nicht befriedigend bewältigt werden, sind die Konsequenzen nicht sehr gravierend. Womöglich erweist sich ein Rendezvous als langweilig oder man ist vom Geschmack eines Mittagessens enttäuscht oder man ist ohne Mantel bei Regenwetter unterwegs. Das wäre alles nicht sehr angenehm. Aber die Tragweite möglicher Konsequenzen bleibt im Treatment B doch weit hinter der Tragweite möglicher Konsequenzen im Treatment A zurück.

Ein Homo Oeconomicus muss unabhängig vom Frame zwingend den Algorithmus gegenüber menschlichen Experten bevorzugen, weil er damit den Erwartungswert seines pekuniären Nutzens maximiert. Jede Entscheidung zugunsten der menschlichen Experten ist als Algorithm Aversion zu betrachten.

Algorithm Aversion ist ein Phänomen, das in unterschiedlichsten Entscheidungssituationen auftreten kann (vgl. Burton, Stein & Jensen, 2020). Folglich gehen wir davon aus, dass das Phänomen auch in dieser Studie zu beobachten ist. Obwohl es anhand der Aufgabenstellung keinen rationalen Grund gibt, auf den menschlichen Experten zurückzugreifen, werden einige Teilnehmer genau dies tun. Hypothese 1 lautet: Es wird nicht ausschließlich der Algorithmus ausgewählt. Nullhypothese 1 lautet daher: Es wird ausschließlich der Algorithmus ausgewählt.

Castelo, Bos & Lehmann (2019) zeigen, dass Framing geeignet ist, die Algorithm Aversion zu beeinflussen. Die Abneigung gegenüber einem Algorithmus ist in verschiedenen Kontexten unterschiedlich stark ausgeprägt. Allerdings ist in dieser Studie der Algorithmus nicht erkennbar die zuverlässigste Alternative und es gibt auch keine erfolgsabhängige Vergütung für die Probanden. Algorithm Aversion wird bei Castelo, Bos & Lehmann (2019) also nicht als Verhaltensanomalie modelliert.

Wir erwarten jedoch, dass es auch dann zu einem Einfluss des Frames auf die Algorithm Aversion kommen wird, wenn die ökonomische Vorteilhaftigkeit des Algorithmus klar erkennbar ist. Hypothese 2 lautet: Der Anteil der Entscheidungen zugunsten des Algorithmus wird sich zwischen den beiden Treatments signifikant unterscheiden. Nullhypothese 2 lautet daher: Der Anteil der Entscheidungen zugunsten des Algorithmus wird sich in beiden Treatments nicht signifikant unterscheiden.

In der Literatur liegen etliche Hinweise darauf vor, dass Framing das Entscheidungsverhalten von Probanden wesentlich beeinflussen kann (vgl. Tversky & Kahneman, 1981). Verhielten sich die Probanden rational und nutzenmaximierend, würde es weder zur Algorithm Aversion noch zum Framing-Effekt kommen. Reale Wirtschaftssubjekte, Menschen aus Fleisch und Blut, verhalten sich jedoch – wie die Forschungsrichtung der Behavioral Economics vielfach belegt – keineswegs wie ein Homo Oeconomicus. Ihr Verhalten entspricht in aller Regel eher dem Modell der Bounded Rationality von Herbert A. Simon (1959). Menschen leiden unter kognitiven Grenzen, sie weichen auf Daumenregeln und Heuristiken aus. Aber sie bemühen sich darum, sinnvolle Entscheidungen zu treffen, sofern dies nicht mit allzu großen Anstrengungen verbunden ist. Eine solche Alltagsvernunft – die gerne beschönigend als „gesunder Menschenverstand" bezeichnet wird – legt nahe, besonders dann große Anstrengungen in Kauf zu nehmen, wenn die Entscheidung besonders gravierende Konsequenzen mit sich bringen kann. Die Gründung eines Unternehmens wird sicherlich in aller Regel gründlicher durchdacht als die Auswahl eines Fernsehprogrammes an einem verregneten Sonntag. Die Auswahl eines Herzchirurgen wird sicherlich in aller Regel mit größerer Sorgfalt betrieben als die Auswahl eines Pizzabringdienstes.

Diese Alltagsvernunft, die unterschiedliche Anstrengungen für Entscheidungssituationen unterschiedlicher Tragweite nahelegt, könnte dazu beitragen, dass die Verhaltensanomalie der Algorithm Aversion in Treatment A (Entscheidungen mit möglicherweise gravierenden Auswirkungen) seltener auftritt als in Treatment B (Entscheidungen mit relativ unbedeutenden Auswirkungen). Hypothese 3 lautet dementsprechend: Je höher die Tragweite einer Entscheidung eingestuft wird, desto seltener tritt die Verhaltensanomalie der Algorithm Aversion auf. Nullhypothese 3 lautet daher: Auch bei steigender Tragweite einer Entscheidungssituation kommt es nicht zu einem Rückgang der Algorithm Aversion.

3
Ergebnisse

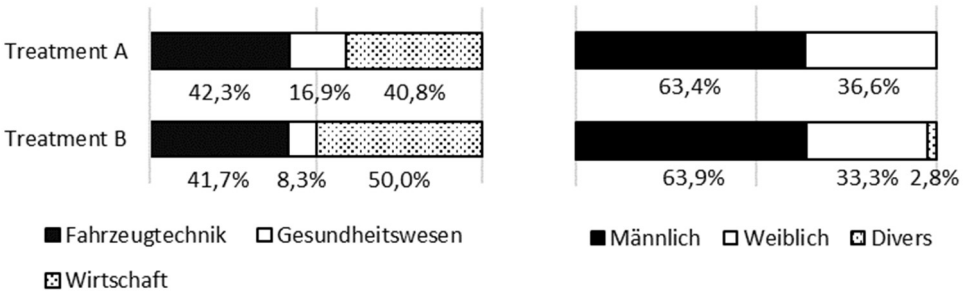
Das ökonomische Experiment findet vom 2. bis zum 14. November 2020 im Ostfalia Labor für experimentelle Wirtschaftsforschung (OLEW) in Wolfsburg statt. Insgesamt nehmen 143 Studierende der Ostfalia Hochschule für angewandte Wissenschaften an dem Experiment teil. Davon sind 91 Probanden (63,6%) männlich, 50 Probanden (35,0%) weiblich und 2 Probanden (1,4%) identifizieren ihre Geschlechtszugehörigkeit als divers. Von den 143 Teilnehmern studieren 65 Probanden (45,5%) an der Fakultät Wirtschaft, 60 Probanden (42,0%) an der Fakultät Fahrzeugtechnik und 18 Probanden (12,6%) an der Fakultät Gesundheitswesen. Das Durchschnittsalter liegt bei 23,5 Jahren.

Von den Teilnehmern befinden sich 71 Probanden in einer Entscheidungssituation, die dem Treatment A zuzuordnen ist. 72 Probanden wird eine Entscheidungssituation präsentiert, die dem Treatment B zuzuordnen ist. Die Aufteilung der Probanden auf die beiden Treatments zeigt sowohl hinsichtlich der Zugehörigkeit zu einer Fakultät als auch im Hinblick auf das Geschlecht ähnliche Verteilungen. In Treatment A (bzw. B) sind 42,3% (41,7%) der Probanden der Fakultät Fahrzeugtechnik zugehörig, 16,9% (8,3%) der Probanden der Fakultät Gesundheitswesen und 40,8% (50,0%) der Fakultät Wirtschaft. In Treatment A (bzw. B) sind 63,4% (63,9%) der Probanden männlich, 36,6% (33,3%) der Probanden weiblich und 0,0% (2,8%) der Probanden divers (Abbildung 1).

Das Experiment ist in z-Tree implementiert (vgl. Fischbacher, 2007). Lediglich die Lotterie zur Ermittlung des Erfolges bei der Durchführung der Dienstleistung wird durch das Ziehen einer Karte aus einem Kartenstapel vorgenommen. Damit wollen wir einem möglichen Verdacht entgegenwirken, dass das Zufallsereignis manipuliert sein könnte. Die Probanden sehen die Spielkarten und können sicher sein, dass bei Wahl des Algorithmus ein Erfolg mit 70%iger Wahrscheinlichkeit (Kartenstapel besteht aus sieben +4€-Karten und drei ±0€-Karten) eintritt. Ferner können sie sicher sein, dass bei Wahl des menschlichen Experten ein Erfolg mit 60%iger Wahrscheinlichkeit (Kartenstapel besteht aus sechs +4€-Karten und vier ±0€-Karten) eintritt (siehe Anhang 4).

Der Zeitaufwand für das Lesen der Spielanleitung (Anhang 1), die Beantwortung der Kontrollfragen (Anhang 2) und die Bearbeitung der Aufgabe liegt bei durchschnittlich 10 Minuten. Eine Antrittsprämie in Höhe von 2,00 € und die Chance auf eine weitere erfolgsabhängige Vergütung in Höhe von 4,00 € erscheint für diesen zeitlichen Aufwand als angemessen. Es sollte sich um einen ausreichenden Anreiz für sinnvolle ökonomische Entscheidungen handeln. Tatsächlich machen die Probanden einen konzentrierten und motivierten Eindruck.

Abbildung 1: Anteile der Probanden hinsichtlich der Zugehörigkeit zu einer Fakultät bzw. des Geschlechts bei Aufteilung auf die Treatments A und B



Die Ergebnisse des Manipulation Checks zeigen, dass die Probanden die Tragweite der Entscheidungssituationen deutlich unterschiedlich wahrnehmen (siehe Tabelle 3 und Abbildung 2). In den gravierenden Entscheidungssituationen (Treatment A) liegt der Mittelwert der wahrgenommenen Tragweite bei einem Wert von 9,00 mit einer Standardabweichung von 1,37. Die Box reicht vom 1. Quartil $x_{0,25} = 8$ (untere Grenze) bis zum 3. Quartil $x_{0,75} = 10$ (obere Grenze). Der Median liegt bei einem Wert von 10. In den Entscheidungssituationen des Treatments B ergibt sich bei der Einschätzung der Tragweite ein Mittelwert von 6,54 bei einer Standardabweichung von 2,53. Die Box reicht vom 1. Quartil $x_{0,25} = 5$ (untere Grenze) bis zum 3. Quartil $x_{0,75} = 8$ (obere Grenze). Der Median liegt bei einem Wert von 7 und somit weit unter dem Median in Treatment A.

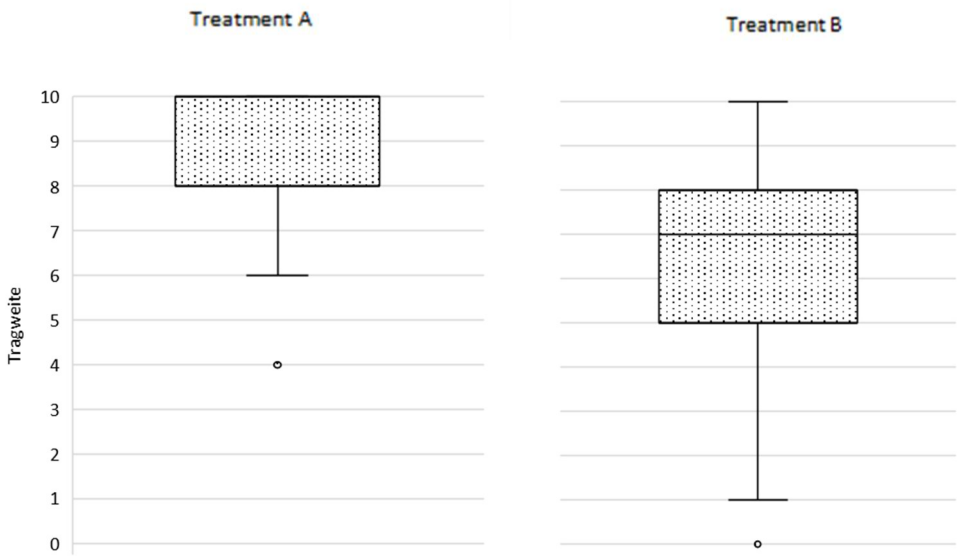
Auch die grafische Analyse zeigt, dass die Probanden die Tragweite in Treatment A insgesamt gravierender einschätzen als in Treatment B. Im direkten Vergleich der Boxplots ist jedoch ersichtlich, dass in Treatment B eine größere Spannweite als in Treatment A vorliegt, da einige Probanden auch in Treatment B die Tragweite als sehr gravierend einschätzen (Abbildung 2).

Tabelle 3: Einschätzung der Tragweite in den Treatments A und B

	Treatment A	Treatment B
Erstes Quartil	8	5
Drittes Quartil	10	8
Median	10	7
Mittelwert	9,00	6,54
Standardabweichung	1,37	2,53

Der Wilcoxon Rank-Sum Test (Mann-Whitney-U-Test) zeigt (vgl. Wilcoxon, 1945; Mann & Whitney, 1947), dass die Tragweite der Entscheidungssituationen in Treatment A signifikant höher als die Tragweite der Entscheidungssituationen in Treatment B eingeschätzt wird (z = 6,689; p = 0,000).

Abbildung 2: Boxplot zur Einschätzung der Tragweite der Entscheidungssituationen



Insgesamt entscheiden sich lediglich 87 von 143 Probanden (60,84%) dazu, die Dienstleistung an den überlegenen Algorithmus zu delegieren. 56 Probanden (39,16%) setzen trotz der niedrigeren Erfolgswahrscheinlichkeit auf menschliche Experten. Nullhypothese 1 muss folglich verworfen werden. Im t-Test ist das Ergebnis hochsignifikant (p = 0,000). Im Durchschnitt unterliegen somit rund zwei von fünf Probanden der Neigung zur Algorithm Aversion (Tabelle 4). Das ist ein überraschend deutliches Ergebnis, da die Entscheidungssituationen sehr übersichtlich sind. Dass eine Bevorzugung menschlicher Experten beziehungsweise die Ablehnung des Algorithmus den Erwartungswert der erfolgsabhängigen Vergütung reduziert, dürfte allen Probanden vollkommen klar sein. Das Bedürfnis, sich dennoch gegen den Algorithmus zu entscheiden, ist bei einem Teil der Probanden offenbar stark ausgeprägt.

Weiterhin lässt sich ein Unterschied in der Anzahl der Entscheidungen zugunsten des Algorithmus zwischen den beiden Treatments beobachten. Während in Treatment A 50,70% der Probanden auf den Algorithmus vertrauen, sind es in Treatment B 70,83% der Probanden. Der χ^2-Test nach Karl Pearson (vgl. Pearson, 1900) offenbart, dass Nullhypothese 2 verworfen werden muss (p =

0,014). Die Häufigkeit, in der Algorithm Aversion auftritt, wird von der Tragweite der Entscheidungssituation beeinflusst. Der Framing-Effekt kommt zum Tragen.

Tabelle 4: Entscheidungen für und gegen den Algorithmus

		Entscheidungen für den Algorithmus		Entscheidungen gegen den Algorithmus	
	n	Anzahl	Prozent	Anzahl	Prozent
Treatment A (gravierend)	71	36	50,70	35	49,30
Treatment B (nicht gravierend)	72	51	70,83	21	29,17
Σ	143	87	60,84	56	39,16

Zwar stellt sich ein Framing-Effekt ein, aber nicht in der vermuteten Art und Weise. Während in Treatment A (mögliche gravierende Konsequenzen) 49,30% der Probanden die Verhaltensanomalie der Algorithm Aversion an den Tag legen, ist dies in Treatment B (keine gravierenden Konsequenzen) nur bei 29,17% der Probanden der Fall (Tabelle 4). Die Nullhypothese 3 kann somit nicht verworfen werden.

Es mag Situationen geben, in denen Menschen sich gerne einmal unvernünftig verhalten. Die Alltagsvernunft legt allerdings nahe, sich diese kleinen Fehltritte dort zu erlauben, wo nicht unmittelbar gravierende Folgen zu befürchten sind. Zum Beispiel sticht der Gastgeber bei einer stimmungsvollen Grillparty noch ein drittes Fass Bier an, obwohl er ahnt, dass das bei einigen seiner Gäste zu Kopfschmerzen am folgenden Tag führen kann. Bei gravierenden Entscheidungen sollte man hingegen hellwach sein und sich von unvernünftigen Neigungen möglichst lösen. Besucht beispielsweise derselbe Mann einen Freund im Krankenhaus, dessen Leben nach einer komplizierten Magenoperation durch Alkoholgenuss akut bedroht wäre, dann verzichtet er besser darauf, ihm eine Flasche seines bevorzugten Whiskys mitzubringen. Diese Gegenüberstellung der beiden Beispiele steht im Einklang mit dem, was man als Alltagsvernunft bezeichnen könnte, und wird die Zustimmung der meisten neutralen Beobachter finden.

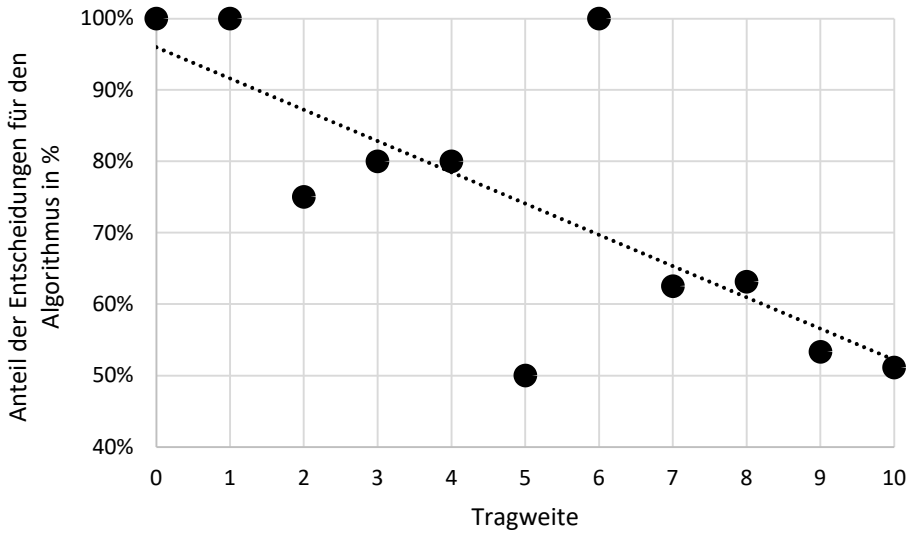
Die Ergebnisse des Experimentes deuten jedoch in die entgegengesetzte Richtung. In den weniger gravierenden Entscheidungssituationen (Treatment B) ist die Neigung zur Algorithm Aversion deutlich geringer ausgeprägt als in den gravierenden Entscheidungssituationen (Treatment A).

Dieses Ergebnis bestätigt sich im Rahmen einer Regressionsanalyse. Dabei wird das Ausmaß der Algorithm Aversion zu der wahrgenommenen Tragweite der

Entscheidungssituation ins Verhältnis gesetzt. Für die möglichen Einschätzungen der Tragweite (von 0 = nicht gravierend bis 10 = sehr gravierend) wird jeweils der dazugehörige durchschnittliche prozentuale Anteil der Entscheidungen zugunsten des Algorithmus ermittelt.

In die Regressionsanalyse gehen die Entscheidungen aller 143 Probanden ein. Die Unterscheidung der beiden Treatments kommt dabei nicht zum Tragen (Abbildung 3).

Abbildung 3: Entscheidungen zugunsten des Algorithmus in Abhängigkeit von der Tragweite der Entscheidungssituation



Käme die oben geschilderte Alltagsvernunft zum Tragen, dann müsste der prozentuale Anteil der Entscheidungen zugunsten des Algorithmus von links nach rechts (also mit steigender wahrgenommener Tragweite der Entscheidungssituation) tendenziell ansteigen. Stattdessen ist das Gegenteil zu beobachten. Während bei einer geringen Tragweite (null und eins) noch 100% der Entscheidungen zugunsten des Algorithmus ausfallen, geht bei zunehmender Tragweite der Anteil der Entscheidungen zugunsten des Algorithmus zurück. Bei einer sehr hohen Tragweite (neun und zehn) entscheidet sich nur noch gut die Hälfte der Probanden dazu, die Dienstleistung von einem Algorithmus ausführen zu lassen (Abbildung 3). Steigt die wahrgenommene Tragweite um eine Einheit, so verringert sich die Wahrscheinlichkeit einer Entscheidung zugunsten des Algorithmus um 3,9% ($t = -2{,}29$; $p = 0{,}023$). Nullhypothese 3 kann somit nicht verworfen werden. In Situationen, die im Falle eines Fehlers schwerwiegende Konsequenzen nach sich ziehen, ist die Algorithm Aversion sogar besonders stark ausgeprägt.

Die Ergebnisse sind sehr überraschend, da die Alltagsvernunft nahe legt, besonders bei Entscheidungen, die schwerwiegende Konsequenzen haben können, die Option mit der höchsten Erfolgswahrscheinlichkeit zu wählen. Wenn sich Wirtschaftssubjekte schon von einer Algorithm Aversion zu ihren eigenen Ungunsten in ihrem Handeln beeinflussen lassen, sollten sie dies eigentlich nur dort tun, wo sie die Konsequenzen guten Gewissens verantworten können. Wo Konsequenzen besonders schwerwiegend sind, müsste die Maximierung der Erfolgsquote Priorität haben. Aber das genaue Gegenteil ist der Fall. Die Algorithm Aversion tritt dort am häufigsten auf, wo sie am meisten Schaden anrichten kann. Insofern drängt es sich auf, von der Tragik der Algorithm Aversion zu sprechen.

Der entscheidende Vorteil eines Framing-Ansatzes besteht darin, dass man in aller Eindeutigkeit den Einfluss eines Faktors herausarbeiten kann. Zwischen den Entscheidungssituationen im Treatment A und den Entscheidungssituationen im Treatment B gibt es nur einen einzigen Unterschied: die Tragweite möglicher Konsequenzen. Freilich handelt es sich dabei um Konsequenzen, die gegebenenfalls von Dritten zu tragen sind. Man könnte die Forschungsanstrengungen fortführen, in dem man den Framing-Ansatz aufgibt und eine unmittelbare Betroffenheit der Probanden modelliert. In diesem Fall muss man unterschiedliche Incentivierungen in den beiden Treatments einführen. Ein Erfolg in Treatment A (gravierende mögliche Konsequenzen) müsste dann mit einem höheren Betrag vergütet werden als im Treatment B (keine gravierenden Konsequenzen). Wir gehen allerdings davon aus, dass sich unsere Ergebnisse auch bei einem solchen experimentellen Ansatz vollumfänglich bestätigen würden. Schließlich handelt es sich um ein Between-Subjects-Design, bei dem jeder Proband nur eine einzige der sechs Entscheidungssituationen vorgelegt bekommt. Ob man nun 4,00 € oder 8,00 € für einen Erfolg im Treatment A erhält, dürfte kaum einen nennenswerten Einfluss auf die Ergebnisse haben. Allerdings bleibt es nachfolgenden Forschungsanstrengungen vorbehalten, diese Einschätzung empirisch zu überprüfen.

4
Zusammenfassung

Viele Menschen entscheiden sich selbst dann gegen den Einsatz eines Algorithmus, wenn klar erkennbar ist, dass der Algorithmus eine höhere Erfolgswahrscheinlichkeit aufweist als ein menschlicher Akteur. Diese Verhaltensanomalie wird als Algorithm Aversion bezeichnet.

Die Probanden nehmen die Perspektive eines Unternehmers ein, der vor der Wahl steht, eine Dienstleistung von einem Algorithmus oder von menschlichen Experten ausführen zu lassen. Wenn die Dienstleistung erfolgreich ausgeführt wird, bekommt der Proband eine erfolgsabhängige Vergütung. Die Probanden werden darüber informiert, dass der jeweilige Algorithmus in 70% aller Fälle zum Erfolg führt, während den menschlichen Experten dies nur in 60% der Fälle gelingt. Angesichts der erkennbar höheren Erfolgsquote spricht alles dafür, auf den Algorithmus zu vertrauen. Dennoch entscheiden sich knapp 40% der Probanden gegen den Algorithmus und für die menschlichen Experten. Damit reduzieren sie den Erwartungswert ihrer erfolgsabhängigen Vergütung und das heißt, sie offenbaren die Verhaltensanomalie der Algorithm Aversion.

Wichtigstes Ziel der Studie ist es herauszufinden, ob Entscheidungssituationen unterschiedlicher Tragweite eine unterschiedliche Häufigkeit für das Auftreten einer Algorithm Aversion mit sich bringen. Dazu wählen wir einen Framing-Ansatz. In sechs Entscheidungssituationen (drei davon mit potenziell gravierenden Auswirkungen und drei davon mit nicht sehr schwerwiegenden möglichen Konsequenzen) besteht eine vollkommen identische Vergütungsstruktur. Die unterschiedliche Tragweite der Entscheidungssituation betrifft nicht die Probanden, sondern gegebenenfalls nur Dritte. Vor diesem Hintergrund bestehen keinerlei Anreize oder Veranlassungen, sich in den sechs Entscheidungssituationen unterschiedlich zu verhalten. Es handelt sich um einen Between-Subjects-Ansatz, das heißt, dass jedem Probanden nur eine der sechs Entscheidungssituationen vorgelegt wird.

Die Ergebnisse fallen deutlich aus. Bei den drei Entscheidungssituationen mit potenziell gravierenden Konsequenzen für Dritte (Treatment A) zeigen knapp 50% der Probanden eine Algorithm Aversion. Bei den drei Entscheidungssituationen mit nicht sehr schwerwiegenden möglichen Konsequenzen für Dritte (Treatment B) zeigen hingegen nur knapp 30% der Probanden eine Algorithm Aversion.

Das ist ein überaus überraschendes Ergebnis. Wenn schon ein Framing-Effekt auftritt, dann hätte man ihn doch eher in umgekehrter Richtung erwartet. Sofern es um Freiheit oder sogar um Kopf und Kragen geht (Treatment A), sollte man doch wohl eher den erfolgsträchtigeren Algorithmus wählen. Stattdessen tritt die Algorithm Aversion gerade hier besonders stark zutage. Wenn es nur um die Anbahnung einer Bekanntschaft, um Wetterprognosen oder um Koch-

rezepte geht (Treatment B), sind die möglichen Konsequenzen recht überschaubar. In einer solchen Situation kann man sich unsinnige Vorbehalte gegenüber einem Algorithmus eher leisten. Aber überraschenderweise tritt in diesen Situationen die Algorithm Aversion verhältnismäßig selten auf.

Man kann es als die Tragik der Algorithm Aversion bezeichnen, dass sie vor allem in Situationen auftritt, in denen sie besonders großen Schaden anrichten kann.

5
Literatur

Alexander, V., Blinder, C. & Zak, P. J. (2018). Why trust an algorithm? Performance, cognition, and neurophysiology, *Computers in Human Behavior*, 89(2018), 279-288.

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2020). Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn, *Business & Information Systems Engineering*, 1-14.

Brozovsky, L. & Petříček, V. (2007). Recommender System for Online Dating Service, ArXiv, abs/cs/0703042.

Burton, J., Stein, M. & Jensen, T. (2020). A Systematic Review of Algorithm Aversion in Augmented Decision Making, *Journal of Behavioral Decision Making*, 33(2), 220-239.

Castelo, N., Bos, M. W. & Lehmann, D. R. (2019). Task-dependent algorithm aversion, *Journal of Marketing Research*, 56(5), 809-825.

Commerford, B. P., Dennis, S. A., Joe, J. R., & Wang, J. (2019). Complex estimates and auditor reliance on artificial intelligence, http://dx.doi.org/10.2139/ssrn.3422591.

Cornelissen, J. & Werner, M. D. (2014). Putting Framing in Perspective: A Review of Framing and Frame Analysis across the Management and Organizational Literature, *The Academy of Management Annals*, 8(1), 181-235.

Dawes, R., Faust, D. & Meehl, P. (1989). Clinical versus actuarial judgment, *Science*, 243(4899), 1668-1674.

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Paper 509, 1-12.

Dietvorst, B. J., Simmons, J. P. & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them, *Management Science*, 64(3), 1155-1170.

Dietvorst, B. J., Simmons, J. P. & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err, *Journal of Experimental Psychology: General*, 144(1), 114-126.

Efendić, E., Van de Calseyde, P. P. & Evans, A. M. (2020). Slow response times undermine trust in algorithmic (but not human) predictions, *Organizational Behavior and Human Decision Processes*, 157(C), 103-114.

Erlei, A., Nekdem, F., Meub, L., Anand, A. & Gadiraju, U. (2020). Impact of Algorithmic Decision Making on Human Behavior: Evidence from Ultimatum Bargaining, *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1), 43-52.

Filiz, I., Judek, J. R., Lorenz, M. & Spiwoks, M. (2021). Reducing Algorithm Aversion through Experience, *Wolfsburg Working Papers*, 21-01.

Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-made Economic Experiments, *Experimental Economics*, 10(2), 171-178.

Frey, B. S. (1992). Behavioural Anomalies and Economics, in: *Economics As a Science of Human Behaviour*, 171-195.

Germann, M., & Merkle, C. (2019). Algorithm Aversion in Financial Investing, http://dx.doi.org/10.2139/ssrn.3364850.

Horne, B. D., Nevo, D., O'Donovan, J., Cho, J. & Adali, S. (2019). Rating Reliability and Bias in News Articles: Does AI Assistance Help Everyone? *ArXiv*, abs/1904.01531.

Ireland, L. (2020). Who errs? Algorithm aversion, the source of judicial error, and public support for self-help behaviors, *Journal of Crime and Justice*, 43(2), 174-192.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards Algorithms? A comprehensive literature Review on Algorithm aversion, *Proceedings of the 28th European Conference on Information Systems (ECIS)*, https://aisel.aisnet.org/ecis2020_rp/168.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk, *Econometrica* 47(2), 263-291.

Kawaguchi, K. (2020). When Will Workers Follow an Algorithm? A Field Experiment with a Retail Business, *Management Science, Articles in Advance*, https://doi.org/10.1287/mnsc.2020.3599.

Köbis, N. & Mossink, L. D. (2020). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry, *Computers in Human Behavior*, 114(2021), 1-13.

Ku, C. Y. (2020). When AIs Say Yes and I Say No: On the Tension between AI's Decision and Human's Decision from the Epistemological Perspectives, *Információs Társadalom*, 19(4), 61-76.

Leyer, M., & Schneider, S. (2019). Me, You or Ai? How Do We Feel About Delegation, *Proceedings of the 27th European Conference on Information Systems (ECIS)*, 1-17.

Logg, J., Minson, J. & Moore, D. (2019). Algorithm appreciation: People prefer algorithmic to human judgment, *Organizational Behavior and Human Decision Processes*, 151 (C), 90-103.

Majumdar, A. & Ward, R. (2011). An algorithm for sparse MRI reconstruction by Schatten p-norm minimization, *Magnetic resonance imaging*, 29(3), 408-417.

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other, *Annals of Mathematical Statistics*, 18(1), 50-60.

Mill, J. S. (1836). On the definition and method of political economy, *The philosophy of economics*, 41-58.

Niszczota, P. & Kaszás, D. (2020). Robo-investment aversion, *PLoS ONE*, 15(9), 1-19.

Önkal, D., Gönül, M. S., & De Baets, S. (2019). Trusting forecasts, *Futures & Foresight Science*, 1(3-4), 1-10.

Pearson, K. (1900). On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157-175.

Persky, J. (1995). The Ethology of Homo Economicus, *Journal of Economic Perspectives*, 9(2), 221-231.

Prahl, A. & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6), 691-702.

Rühr, A., Streich, D., Berger, B. & Hess, T. (2019). A Classification of Decision Automation and Delegation in Digital Investment Systems, *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 1435-1444.

Sawaitul, S. D., Wagh, K. & Chatur, P.N. (2012). Classification and Prediction of Future Weather by using Back Propagation Algorithm-An Approach, *International Journal of Emerging Technology and Advanced Engineering*, 2(1), 110-113.

Shariff, A., Bonnefon, J. F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles, *Nature Human Behaviour*, 1(10), 694-696.

Simon, H. A. (1959). Theories of Decision-Making in Economics and Behavioral Science, *The American Economics Review*, 49(3), 253-283.

Simpson, B. (2016). Algorithms or advocacy: does the legal profession have a future in a digital world? *Information & Communications Technology Law*, 25(1), 50-61.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice, *Science*, 211(4481), 453-458.

Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases, *Science*, 185(4157), 1124-1131.

Ueda, M., Takahata, M. & Nakajima, S. (2011). User's food preference extraction for personalized cooking recipe recommendation, *Proceedings of the Second International Conference on Semantic Personalized Information Management: Retrieval and Recommendation*, 781, 98-105.

Wang, R., Harper, F. M., & Zhu, H. (2020, April). Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Paper 684, 1-14.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods, *Biometrics Bulletin*, 1(6), 80-83.

Yeomans, M., Shah, A. K., Mullainathan, S. & Kleinberg, J. (2019). Making Sense of Recommendations, *Journal of Behavioral Decision Making*, 32(4), 403-414.

6
Anhang

6.1
Anhang 1: Spielanleitung

Das Spiel

Sie sind Unternehmer/-in und müssen entscheiden, ob Sie eine erstmalig von Ihnen angebotene Dienstleistung ausschließlich von einem Algorithmus oder ausschließlich von menschlichen Experten durchführen lassen. Es ist bekannt, dass die menschlichen Experten die Aufgabe mit einer Wahrscheinlichkeit von 60% erfolgreich abschließen. Weiterhin ist bekannt, dass der Algorithmus die Aufgabe mit einer Wahrscheinlichkeit von 70% erfolgreich abschließt.

Ablauf

Nach dem Lesen der Instruktionen und der Beantwortung der Kontrollfragen wird Ihnen die Entscheidungssituation präsentiert. Daraus geht hervor, welche Dienstleistung Ihr Unternehmen konkret anbietet. Zunächst werden Sie gebeten, die Tragweite der Entscheidungssituation aus Sicht Ihrer Kunden zu beurteilen. Anschließend treffen Sie die Entscheidung, ob die Dienstleistung von menschlichen Experten oder von einem Algorithmus durchgeführt werden soll.

Auszahlung

Für die Teilnahme am Experiment erhalten Sie eine Antrittsprämie von 2,00€. Außerdem erhalten Sie eine Bonuszahlung von weiteren 4,00€, wenn die Dienstleistung erfolgreich durchgeführt wird.

Hinweise

- Bitte verhalten Sie sich während des Experiments ruhig!
- Bitte schauen Sie Ihren Nachbarn nicht auf den Bildschirm!
- Bis auf einen Stift und einen Taschenrechner sind **keine** Hilfsmittel (Smartphones, Smartwatches etc.) zugelassen.

6.2
Anhang 2: Kontrollfragen

Kontrollfrage 1:

Welche Alternativen stehen Ihnen für Ihre Dienstleistung zur Verfügung?

- Ich kann die Dienstleistung selbst durchführen oder von einem Algorithmus ausführen lassen.
- Ich kann die Dienstleistung selbst durchführen oder von menschlichen Experten ausführen lassen.
- Ich kann die Dienstleistung von menschlichen Experten oder von einem Algorithmus ausführen lassen. (Zutreffend!)

Kontrollfrage 2:

Für wie viele neu angebotene Dienstleistungen treffen Sie Ihre Auswahl?

- Keine.
- Eine. (Zutreffend!)
- Zwei.

Kontrollfrage 3:

Wie hoch ist die Bonuszahlung für eine erfolgreiche Durchführung?

- 1,00 €.
- 2,50 €.
- 4,00 €. (Zutreffend!)

Kontrollfrage 4:

Wie hoch ist die Bonuszahlung für eine fehlerhafte Durchführung?

- -2,50 €.
- 0,00 €. (Zutreffend!)
- 2,50 €.

6.3
Anhang 3: Entscheidungssituationen Treatment A und B

Treatment A: Eher gravierende Entscheidungssituationen

Entscheidungssituation A-1: Autonomes Fahren

Sie sind Leiter/-in eines Beförderungsunternehmens und müssen entscheiden, ob Sie Ihre 100.000 Kunden ausschließlich mit autonomen Fahrzeugen (Algorithmus) oder ausschließlich mit Fahrzeugen mit Chauffeur (menschliche Experten) an ihr Ziel bringen. Die Aufgabe gilt als erfolgreich gelöst, wenn alle Kunden sicher ans Ziel gebracht wurden. Eine falsche Entscheidung kann im Extremfall den Tod eines Passagiers bedeuten.

Ich wähle: O Autonome Fahrzeuge (Algorithmus)
O Chauffeure (menschliche Experten)

Entscheidungssituation A-2: MRT-Scan

Sie sind Leiter/in eines großen Krankenhauses und müssen entscheiden, ob Sie die MRT-Scans Ihrer 100.000 Patienten mit Hirnkrankheiten ausschließlich von einem spezialisierten Computerprogramm (Algorithmus) oder ausschließlich von Ärzten (menschliche Experten) auswerten lassen. Die Aufgabe gilt als erfolgreich gelöst, wenn sämtliche lebensbedrohliche Symptome unmittelbar erkannt werden. Eine falsche Entscheidung kann im Extremfall den Tod eines Patienten bedeuten.

Ich wähle: O Spezialisiertes Computerprogramm (Algorithmus)
O Ärzte (menschliche Experten)

Entscheidungssituation A-3: Strafverfahren

Sie sind Leiter/-in einer großen Anwaltskanzlei und müssen entscheiden, ob Sie die Analyse der Verfahrensunterlagen Ihrer 100.000 Mandanten ausschließlich von einem spezialisierten Computerprogramm (Algorithmus) oder ausschließlich von Strafverteidigern (menschliche Experten) durchführen lassen. Die Aufgabe gilt als erfolgreich gelöst, wenn der Mittelwert der Strafen Ihrer Mandanten unter dem Bundesdurchschnitt liegt. Eine falsche Entscheidung kann für einen Mandanten im Extremfall eine ungerechtfertigte jahrelange Freiheitsstrafe bedeuten.

Ich wähle: O Spezialisiertes Computerprogramm (Algorithmus)
O Strafverteidiger (menschliche Experten)

Treatment B: Eher weniger gravierende Entscheidungssituationen

Entscheidungssituation B-1: Partnervermittlung

Sie sind Leiter/-in einer Online-Partnervermittlung und müssen entscheiden, ob Sie Ihre 100.000 Kunden ausschließlich von einem spezialisierten Computerprogramm (Algorithmus) oder ausschließlich von geschulten Mitarbeitern (menschliche Experten) an andere Nutzer ihrer Plattform vermitteln lassen. Die Aufgabe gilt als erfolgreich gelöst, wenn Sie die Bewertung Ihrer App im App-Store verbessern können. Eine falsche Entscheidung kann für ihre Kunden zu einem Date mit einem nicht optimal ausgewählten Kandidaten führen.

Ich wähle: O Spezialisiertes Computerprogramm (Algorithmus)
O Geschulte Mitarbeiter (menschliche Experten)

Entscheidungssituation B-2: Kochrezepte

Sie sind Leiter/-in eines Online-Lebensmittelhändlers und müssen entscheiden, ob Sie Ihre 100.000 Kochboxen mit individuell auf den Kunden abgestimmten Zutaten und Rezepten ausschließlich von einem spezialisierten Computerprogramm (Algorithmus) oder ausschließlich von geschulten Mitarbeitern (menschliche Experten) zusammenstellen lassen. Die Aufgabe gilt als erfolgreich gelöst, wenn Sie die Wiederbestellquote erhöhen können, die als maßgeblicher Indikator der Kundenzufriedenheit gilt. Eine falsche Entscheidung kann dazu führen, dass den Kunden ihr Essen nicht schmeckt.

Ich wähle: O Spezialisiertes Computerprogramm (Algorithmus)
O Geschulte Mitarbeiter (menschliche Experten)

Entscheidungssituation B-3: Wetterprognose

Sie sind Leiter/-in eines Nachrichtenportals und müssen entscheiden, ob Sie Ihre 100.000 täglichen Wetterprognosen für verschiedene Städte ausschließlich von einem spezialisierten Computerprogramm (Algorithmus) oder ausschließlich von erfahrenen Wissenschaftlern (menschliche Experten) durchführen lassen. Die Aufgabe gilt als erfolgreich gelöst, wenn die am Vortag prognostizierte Temperatur im Schnitt nicht mehr als 1 Grad von der tatsächlichen Temperatur abweicht. Eine falsche Entscheidung kann dazu führen, dass sich Ihre Leser nicht dem Wetter entsprechend anziehen.

Ich wähle: O Spezialisiertes Computerprogramm (Algorithmus)
O Erfahrene Wissenschaftler (menschliche Experten)

6.4
Anhang 4: Ermittlung des Zufallsereignisses mit Hilfe einer Lotterie

Abbildung 4: Kartenstapel bei Wahl des Algorithmus

Kartenstapel bei Wahl des Algorithmus: Sieben Karten mit dem Ereignis +4€ und drei Karten mit dem Ereignis ±0€.

Abbildung 5: Kartenstapel bei Wahl des menschlichen Experten

Kartenstapel bei Wahl des menschlichen Experten: Sechs Karten mit dem Ereignis +4€ und vier Karten mit dem Ereignis ±0€.

6.5
Anhang 5: Definitionen Algorithm Aversion in der Literatur

Autoren	Definition „Algorithm Aversion“ (englisch)
Dietvorst, Simmons & Massey, 2015	"Research shows that evidence-based algorithms more accurately predict the future than do human forecasters. Yet when forecasters are deciding whether to use a human forecaster or a statistical algorithm, they often choose the human forecaster. This phenomenon, which we call *algorithm aversion (…)*"
Prahl & Van Swol, 2017	"The irrational discounting of automation advice has long been known and a source of the spirited “clinical versus actuarial” debate in clinical psychology research (Dawes, 1979; Meehl, 1954). Recently, this effect has been noted in forecasting research (Önkal et al., 2009) and has been called algorithm aversion (Dietvorst, Simmons, & Massey, 2015)."
Dietvorst, Simmons & Massey, 2018	"Although evidence-based algorithms consistently outperform human forecasters, people often fail to use them after learning that they are imperfect, a phenomenon known as *algorithm aversion*."
Castelo, Bos & Lehmann, 2019	"The rise of algorithms means that consumers are increasingly presented with a novel choice: should they rely more on humans or on algorithms? Research suggests that the default option in this choice is to rely on humans, even when doing so results in objectively worse outcomes."
Commerford, Dennis, Joe & Wang, 2019	“(…) *algorithm aversion* – the tendency for individuals to discount computer-based advice more heavily than human advice, although the advice is identical otherwise.”
Horne, Nevo, O’Donovan, Cho & Adali, 2019	“For example, Dietvorst et al. (Dietvorst, Simmons, and Massey 2015) studied when humans choose the human forecaster over a statistical algorithm. The authors found that aversion of the automated tool increased as humans saw the algorithm perform, even if that algorithm had been shown to perform significantly better than the human. Dietvorst et al. explained that aversion occurs due to a quicker decrease in confidence in algorithmic forecasters over human forecasters when seeing the same mistake occur (Dietvorst, Simmons, and Massey 2015).”
Ku, 2019	“(…) “algorithm aversion”, a term refers by Dietvorst et al. (Dietvorst et al. 2015) means that humans distrust algorithm even though algorithm consistently outperform humans.”
Leyer & Schneider, 2019	“In the particular context of the delegation of decisions to AI-enabled systems, recent findings have revealed a general algorithmic aversion, an irrational discounting of such systems as suitable decision-makers despite objective evidence (Dietvorst, Simmons and Massey, 2018)”
Logg, Minson & Moore, 2019	"(…) human distrust of algorithmic output, sometimes referred to as “algorithm aversion” (Dietvorst, Simmons, & Massey, 2015).[1] “; Fußnote 1: "while this influential paper [of Dietvorst et al.] is about the effect that seeing an algorithm err has on people's likelihood of choosing it, it has been cited as being about how often people use algorithms in general."
Önkal, Gönül & De Baets, 2019	“(…) people are averse to using advice from algorithms and are unforgiving toward any errors made by the algorithm (Dietvorst et al., 2015; Prahl & Van Swol, 2017).”
Rühr, Streich, Berger & Hess, 2019	"Users have been shown to display an aversion to algorithmic decision systems [Dietvorst, Simmons, Massey, 2015] as well as to the perceived loss of control associated with excessive delegation of decision authority [Dietvorst, Simmons, Massey, 2018]."
Yeomans, Shah, Mullainathan & Kleinberg, 2019	"(...) people would rather receive recommendations from a human than from a recommender system (...). This echoes decades of research showing that people are averse to relying on algorithms, in which the primary driver of aversion is algorithmic errors (for a review, see Dietvorst, Simmons, & Massey, 2015)."

Berger, Adam, Rühr & Benlian, 2020	“Yet, previous research indicates that people often prefer human support to support by an IT system, even if the latter provides superior performance – a phenomenon called algorithm aversion.” (…) “These differences result in two varying understandings of what algorithm aversion is: unwillingness to rely on an algorithm that a user has experienced to err versus general resistance to algorithmic judgement.”
Burton, Stein & Jensen, 2020	"(…) algorithm aversion—the reluctance of human forecasters to use superior but imperfect algorithms— (…)"
De-Arteaga, Fogliato & Chouldechova, 2020	“*Algorithm aversion*–the tendency to ignore tool recommendations after seeing that they can be erroneous (…)”
Efendić, Van de Calseyde & Evans, 2020	"Algorithms consistently perform well on various prediction tasks, but people often mistrust their advice."; “However, repeated observations show that people profoundly mistrust algorithm-generated advice, especially after seeing the algorithm fail (Bigman & Gray, 2018; Diab, Pui, Yankelevich, & Highhouse, 2011; Dietvorst, Simmons, & Massey, 2015; Önkal, Goodwin, Thomson, Gönül, & Pollock, 2009).”
Erlei, Nekdem, Meub, Anand & Gadiraju, 2020	"Recently, the concept of algorithm aversion has raised a lot of interest (see (Burton, Stein, and Jensen 2020) for a review). In their seminal paper, (Dietvorst, Simmons, and Massey 2015) illustrate that human actors learn differently from observing mistakes by an algorithm in comparison to mistakes by humans. In particular, even participants who directly observed an algorithm outperform a human were less likely to use the model after observing its imperfections."
Germann & Merkle, 2020	“The tendency of humans to shy away from using algorithms even when algorithms observably outperform their human counterpart has been referred to as algorithm aversion.”
Ireland, 2020	“(…) some researchers find that, when compared to humans, people are averse to algorithms after recording equivalent errors.”
Jussupow, Benbasat & Heinzl, 2020	"(…) literature suggests that although algorithms are often superior in performance, users are reluctant to interact with algorithms instead of human agents – a phenomenon known as algorithm aversion"
Kawaguchi, 2020	“The phenomenon in which people often obey inferior human decisions, even if they understand that algorithmic decisions outperform them, is widely observed. This is known as algorithm aversion (Dietvorst et al. 2015).”
Köbis & Mossink, 2020	“When people are informed about algorithmic presence, extensive research reveals that people are generally averse towards algorithmic decision makers. This reluctance of “human decision makers to use superior but imperfect algorithms” (Burton, Stein, & Jensen, 2019; p.1) has been referred to as algorithm aversion (Dietvorst, Simmons, & Massey, 2015). In part driven by the belief that human errors are random, while algorithmic errors are systematic (Highhouse, 2008), people have shown resistance towards algorithms in various domains (see for a systematic literature review, Burton et al., 2019).”
Niszczota & Kaszás, 2020	“When given the possibility to choose between advice provided by a human or an algorithm, people show a preference for the former and thus exhibit algorithm aversion (Castelo et al., 2019; Dietvorst et al., 2015, 2016; Longoni et al., 2019).”
Wang, Harper & Zhu, 2020	“(…) people tend to trust humans more than algorithms even when the algorithm makes more accurate predictions.”