LBBW Bereit für Neues

Kapitalmärkte Blickpunkt

Ausgabe 24.11.2023 I LBBW Research I Macro/Strategy

Generative KI hat hohes Diskriminierungspotenzial

Auf einen Blick

- In Software spiegeln sich die kulturellen Werte ihrer Entwickler wider.
- Die Berücksichtigung ethischer Prinzipien bei der Modellentwicklung sind daher grundlegend für eine nachhaltige Nutzung von künstlicher Intelligenz.
- Generative KI wirft neue ethische Fragen auf. Es entstehen neue Risiken, Frauen und marginalisierte Gruppen zu diskriminieren.
- Die Entwicklung einer KI-Ethik ist daher essenziell bei der Entwicklung einer KI-Strategie und der Anwendung von KI-Modellen durch Unternehmen.

Dr. Guido Zimmermann
Senior Economist
+49 711 127-71640
Guido.Zimmermann@LBBW.de

LBBWResearch@LBBW.de
LBBW_Research

Erstellt am:
24.11.2023 09:39

Ethische Aspekte generativer KI

Software spiegelt die Werte ihrer Entwickler, deren Unternehmens und eines Landes wider. Die USA haben andere Maßstäbe als beispielsweise China, die EU oder Deutschland. Die Werte eines Landes unterscheiden sich auch dadurch, welche Bedeutung sie den Rechten der Bevölkerung und insbesondere denen von marginalisierten Gruppen beimisst. Da Software immer mehr unser Leben bestimmt, ist es grundlegend, eine gemeinsame Vorstellung von ethischen Prinzipien zu entwickeln, die für die Entwicklung und Anwendung von Applikationen künstlicher Intelligenz (KI) gelten. Zum einen aus fundamentalen gesellschaftlichen Gründen. Zum anderen aber auch, weil Anbieter und Nutzer unter Umständen ihre Reputation aufs Spiel setzen, wenn sie keine ausreichenden moralischen Maßstäbe anlegen.

> Software spiegelt kulturelle Werte wider

Generative KI (GenAI – Generative Artificial Intelligence) ist wahrscheinlich eine neue Basistechnologie, die sehr viele Bereiche der Gesellschaft berühren und folglich verändern wird. Generative KI ist deswegen so bedeutsam, weil zum einen ihre Nutzer selbst sehr schnell digitalen Inhalt kreieren können. Und zum anderen, weil praktisch jedes Problem, das

einen Prognosecharakter und einen Bezug zu Sprache hat, perspektivisch mit Methoden generativer KI gelöst werden kann.

Wichtig ist zu verstehen, dass durch GenAI möglicherweise eine neue Ära des Computing entsteht. In der Vergangenheit funktionierte ein Computer quasi wie ein Taschenrechner: Er spuckte auf eine Anfrage ein immer gleiches, korrektes Ergebnis aus. Durch GenAI nimmt Computing statt eines deterministischen nun einen probabilistischen Charakter an. Denn generative KI funktioniert nicht wie ein Taschenrechner, sondern wie ein Wahrscheinlichkeitsrechner. Sie berechnet auf Grundlage von Daten etwa bei Textanwendungen die Wahrscheinlichkeit, mit der Worte aufeinander folgen. GenAI-Modelle gehen daher mit einer gewissen Fehlerquote einher. Diese Fehlerquote ist bei traditionellen KI-Modellen viel geringer. Die Fehlerquote von GenAI-Modellen nennt man „Halluzinationen“: Ein GenAI-Modell produziert zu einem gewissen Grad plausibel klingende, aber nicht korrekte Antworten in Reaktion auf Eingaben der Anwender. Fehler dürfen sich aber gerade Institutionen nicht erlauben.

Drei Dinge sind auf absehbare Zeit in Bezug auf GenAI wohl nicht lösbar:

- Das Auftreten von Halluzinationen: plausibel klingenden, aber falschen Antworten.
- Bislang existieren keine statistischen Gütekriterien für GenAI-Modelle, die vergleichbar sind mit denen bei klassischen KI-Modellen.
- GenAI-Modelle basieren auf neuronalen Netzen. Sie stellen quasi eine „Black Box“ dar. Die Entscheidungen eines KI-Modells können bislang nicht erklärt bzw. interpretiert werden.

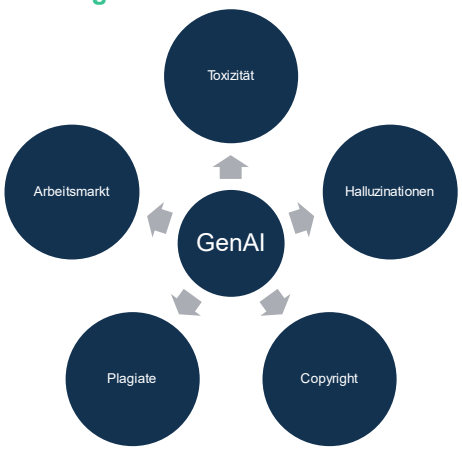
Risiken von GenAI

GenAI-Modelle gehen mit neuen ethischen Herausforderungen einher:

- Selbst von führenden KI-Forschern wird es für möglich gehalten, dass eine KI-Superintelligenz entstehen könnte, die im schlimmsten Fall in der Lage ist, die Menschheit auszulöschen. Ausgeschlossen werden kann ein derartiges Extremrisiko natürlich nicht. Wir sehen es aber nicht als das Hauptrisiko an.
- Ein viel größeres Risiko ist eine Disruption des Arbeitsmarkts für Berufe mit einem überwiegenden Anteil kognitiver Tätigkeiten. Langfristig sehen wir darin kein Problem. Kurz- bis mittelfristig können hier aber Probleme auf dem Arbeitsmarkt entstehen.
- GenAI-Modellen können durch Kriminelle und feindliche (staatliche) Akteure missbraucht werden. Zu denken ist hier an Betrug, Erpressung und die Beeinflussung von Wahlen durch KI-generierte Fehlinformationen.
- Menschen, die keinen Zugang zu GenAI-Modellen haben, werden von dieser neuen Technologie ausgeschlossen.
- Die GenAI-Modellen zugrundeliegenden großen Sprachmodelle sind bislang wenig robust. Die Qualität ihrer Ergebnisse differiert stark. Das Problem der Halluzinationen ist bislang nicht gelöst.
- GenAI-Modelle gehen mit großen Problemen des Datenschutzes, neuen Cyberrisiken und Copyright-Problemen einher.
- GenAI-Modelle weisen Probleme des Bias und der Diskriminierung gegenüber und von Frauen und marginalisierten Gruppen auf. Unter Bias versteht man, dass ein Modell systematisch eine Schlagseite hat und beispielsweise Frauen systematisch gegenüber Männern diskriminiert.

- Die bislang populären GenAI-Modelle stammen in erster Linie von Anbietern aus den USA. Es ist noch nicht geklärt, inwieweit diese Modelle den Werten europäischer und deutscher Unternehmen entsprechen. GPT-Modelle teilen größtenteils die Werte gebildeter Schichten in den westlichen Industrieländern. Erstaunlicherweise liegen die dort implizierten Werte sehr nahe an den Werten Deutschlands, spiegeln aber nicht die Werte von vielen Schwellenländern oder konservativeren Ländern Europas wider.
- Zusätzlich weisen die Modelle oft einen erheblichen CO_2-Fußabdruck auf. Hinzu kommt ein immenser Wasserverbrauch bei der Kühlung der Datenzentren. Andererseits dürften die Ziele der Nachhaltigkeit ohne KI wohl nicht zu stemmen sein.

Spezifische Probleme generativer KI



Quelle: Amazon Science, LBBW Research

Probleme der Diskriminierung

Wie kann KI diskriminieren? Eine bewusste Diskriminierung durch KI-Modelle, die darauf basiert, dass sie gezielt so programmiert sind, halten wir für einen Ausnahmefall. KI-Modelle basieren aber auf der Analyse und dem Finden von Mustern in Massendaten (Big Data). GenAI-Modelle produzieren Antworten, die repräsentativ für diese Massendaten sind. Die KI-Anbieter sind an der Repräsentativität der Daten interessiert. Spezifische Daten von marginalisierten Gruppen werden nicht adäquat berücksichtigt.

Big Data berücksichtigt nicht Daten von Minderheiten

KI-Modelle können vor allem durch eine statistische Diskriminierung Frauen und marginalisierte Gruppen verletzen oder Stereotype und Vorurteile bestätigen und verstärken. Eine statistische Diskriminierung kann beispielsweise dadurch entstehen, dass ein Kunde, der die Postleitzahl eines unterprivilegierten Wohnviertels oder ein zu niedriges Einkommen aufweist oder nicht einen für die Mehrheit der Bevölkerung repräsentativen typischen Namen trägt, durch KI-Systeme diskriminiert wird, weil man diesem Kunden eine zu geringe Kreditwürdigkeit beimisst. Dieses Beispiel der traditionellen KI ist wohlbekannt und im Prinzip durch entsprechende Anpassungen behebbar. GenAI-Modelle dagegen gehen mit neuen, noch ungelösten Problemen einher. Beispiele und Ursachen für Diskriminierungen und Biases in GenAI-Modellen sind beispielsweise bei der Generierung von Bildern folgende:

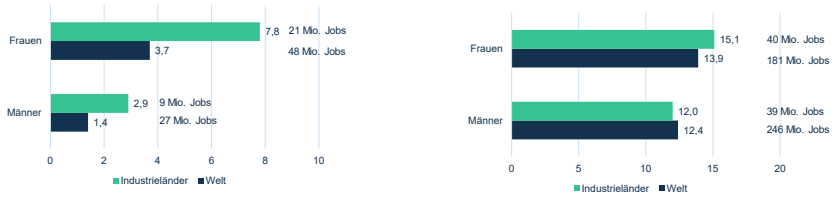
- Die Daten spiegeln vor allem die US-Sicht wider.
- Generierte Bilder zeigen meist Personen, die westlichen Schönheitsstandards entsprechen: junge, kaukasisch aussehende Frauen mit heller Haut und langem Haar. Selbst schwarze Menschen werden typischerweise mit einer helleren Haut dargestellt als es in der Realität der Fall ist.
- Generierte Bilder von Menschen aus Schwellenländern transportieren zumeist Klischees aus westlicher Sicht (wie etwa die Darstellung der Menschen in traditioneller Kleidung oder eine verzerrte, klischeebehaftete Darstellung der Wohnverhältnisse).
- Bei der Darstellung von beruflichen Tätigkeiten werden beispielsweise Schwarze selbst dann nicht als Mediziner dargestellt, wenn das Modell explizit dazu aufgefordert wird. Frauen werden mit „klassischen" Frauenberufen assoziiert (wie etwa Sekretärin und nicht Data Scientist oder Chirurg).

Wie können Frauen durch GenAI-Modelle diskriminiert werden?

Diskriminierung von Frauen

- Bei der Modellentwicklung ist es essenziell, dass Mitarbeiter die KI-Modelle trainieren, indem sie auf die Ergebnisse der KI Feedback geben (Feedback-Loop). Diese stellen nämlich sicher, dass als anstößig oder von den Tech-Unternehmen als gefährlich empfundene Inhalte nicht durch die KI-Modelle ausgewertet werden. Zu denken ist hier etwa an die Beschreibung und Darstellung sexueller und krimineller Inhalte in Text und Bild. Diese relativ stumpfsinnige Arbeit kann nicht nur psychologische Traumata hinterlassen, sondern wird auch zumeist von niedrigbezahlten Mitarbeitern in Schwellenländern oder gar von Gefängnisinsassen durchgeführt. Diese Click-Workers, die auch weiblich sein können, werden hier unter Umständen mehrfach ausgebeutet. Andererseits bieten derartige Stellen neue Verdienstmöglichkeiten für arme Menschen in Schwellenländern.
- In den Daten und Algorithmen versteckte Werturteile können Frauen und marginalisierten Gruppen beispielsweise bei Bewerbungen benachteiligen. GenAI-Modelle verwenden eine sehr männliche Sprache (etwa bei Stellenausschreibungen). GenAI-Modelle sollten daher nicht für Entscheidungen und Werturteile verwendet werden. Eine Antwort auf eine Bewerbung ist ein Werturteil.
- Frauen, die in leicht automatisierbaren Büroberufen tätig sind, können durch die Anwendung von GenAI auf dem Arbeitsmarkt stark getroffen werden. Schlimmstenfalls werden ihre Stellen wegrationalisiert. Frauen sind relativ oft in der Bürosacharbeit tätig. GenAI-Anwendungen sind gerade in Berufen mit solchen Tätigkeitsprofilen stark und können zur Automatisierung führen.
- Frauen sind in den Datenbasen des Tech- und Gesundheitssektors unterrepräsentiert und entsprechend unsichtbar. Ihre spezifischen Bedürfnisse werden von der (medizinischen) Forschung weniger berücksichtigt als die von Männern.
- Es ist schon heute ein Leichtes, pornographische Inhalte unter Verwendung von Fotos zu kreieren. Das Problem von Deep Fake Porn dürfte für Frauen virulent werden.
- Bei der Generierung von Fotos können Stereotype zementiert werden, wenn etwa bei der Frage nach einem Bild für Büroassistenten lediglich Bilder von Frauen generiert werden, bei der Frage nach Bildern von Chirurgen aber nur Bilder von Männern.

Potenziale für Automatisierung (l. S.) und Verstärkung (r. S.) von Tätigkeiten durch KI



Quelle: ILO, LBBW Research

Arbeitsplätze von Frauen häufiger bedroht

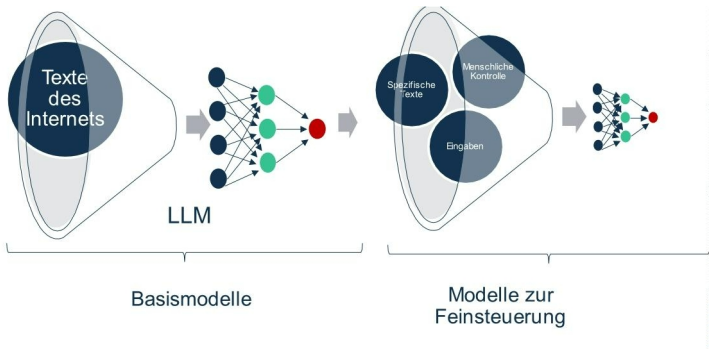
Die in diesen Beispielen zu findenden Diskriminierungspotenziale sind vor allem im Daten-Bias zu suchen:

- Die Schlagseite der Daten resultiert zum einen aus der Gesellschaftsstruktur. KI spiegelt die Vorurteile einer Gesellschaft wider. Der Daten-Bias kann aber wiederum die Gesellschaftsstruktur formen. Beides beeinflusst sich gegenseitig.
- Die Antworten von großen Sprachmodellen entsprechen am ehesten denen von weißen, gebildeten, reichen Menschen aus demokratischen Industrieländern.
- Der Daten-Bias resultiert auch aus der homogenen soziologischen Struktur der Tech-Community: In erster Linie schreiben junge, weiße Männer aus Kalifornien, die zumeist aus dem Bürgertum stammen, die Software. Auch aus dieser Sicht ist KI eine Kulturtechnologie: Die bekanntesten GenAI-Modelle spiegeln die Werte des Silicon Valley wider.
- Neuere Untersuchungen zeigen, dass GenAI-Modelle einen politischen Bias haben können, der sich in der Weltsicht und den Meinungen des Modells äußert. Textgenerator-Modelle sind daher derzeit eine Spielwiese für die Kulturkämpfe in den USA, die über ihre Verwendung nach Europa transferiert werden. Entsprechend wünscht die Regierung Chinas, dass die dort entwickelten Modelle sozialistische Werte widerspiegeln.
- Die Tech-Konzerne sind eher an der Repräsentativität als an spezifischen Merkmalen der Daten interessiert. Die Mehrheitsbevölkerung hat damit naturgemäß eine größere Bedeutung als die Repräsentanz von marginalisierten Gruppen in Daten.

Diskriminierungspotenziale in allen Stufen des Modellbaus

Wo in der Kette des KI-Modellbaus können Diskriminierungspotenziale entstehen? Die folgende stilisierte Darstellung der Konstruktion von großen Sprachmodellen (Large Language Models – LLMs) zeigt auf, dass quasi bei allen Stellschrauben im Modellbau ein Bias entstehen kann – angefangen beim Grunddatenkorpus der Basismodelle (den Texten und Bildern des Internets) bis hin zu den verwendeten Texten der jeweiligen Institution und der menschlichen Kontrolle des Modells, sowie bei Eingaben, die das Modell trainieren.

Möglichkeiten des Bias beim GenAI-Modellbau bei allen Inputs



Quelle: LBBW Research

Entwicklung von KI-Ethik

Die Entwicklung einer KI-Ethik in den Unternehmen ist ein Prozess, der notwendigerweise zur KI-Produktentwicklung gehören muss. Dieser Prozess unterliegt Prinzipien. Die Entwicklung einer KI-Ethik muss zu einer Daten-, Digitalisierungs- und KI-Strategie auch eines Unternehmens gehören, das große Sprachmodelle für seine Zwecke feinsteuert. Welchen Prioritäten hier aber gelten sollen, darüber besteht in den Unternehmen bislang kein Konsens. Welche Fragen müssen bei der Produktentwicklung beantwortet werden?

- Dient das Produkt allen Nutzern in gleichem Maße? Weist die entwickelte Dienstleistung eine Mindestqualität auf? Wer sind die verletzlichsten Gruppen? Können die Metriken, die den Erfolg des Produkts bestimmen, nach Gruppen aufgeteilt werden? Für welche Gruppen weisen die Metriken eine schlechtere Nutzererfahrung auf?
- Werden alle gesellschaftlichen Gruppen gleich behandelt (etwa über Sprachen und Weltregionen hinweg)?
- Gibt es Nutzergruppen, für die das Produktergebnis besondere Beachtung verdient (beispielsweise bei einem historischen Kontext der Inhalte)?

Was muss in Zukunft bedacht und entwickelt werden?

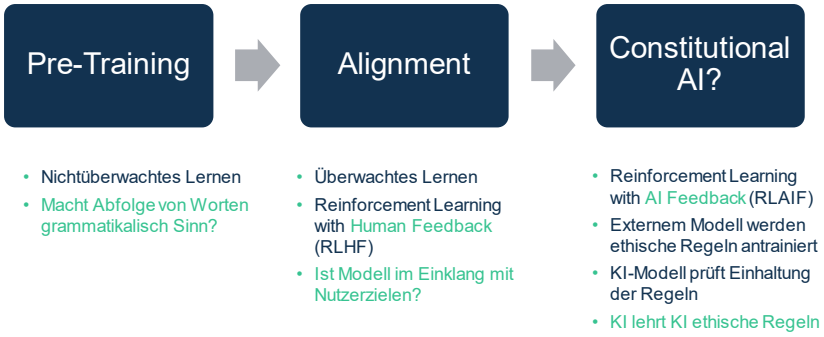
- Die Sichtbarkeit von Frauen und marginalisierten Gruppen muss im Tech-Sektor und in den Datensets erhöht werden. Dies dürfte ein langwieriger Prozess werden.
- Die Unternehmen sollten die neuesten verfügbaren Modelle verwenden, die perspektivisch etwaige Biases nicht mehr in sich tragen. Damit einher gehen sollten sogenannte GPT-Wetterprognosen, die den Anwendern täglich aufzeigen, wie die aktuelle Modellperformance ist. In neueren Modelltypen dürften nämlich perspektivisch Menschen bei der Feinsteuerung durch KI-Systeme ersetzt werden, die automatisch die KI-Modellentwicklungen dahingehend überprüfen, ob sie beispielsweise der UN-Menschenrechtskonvention, den KI-Prinzipien der OECD, dem Grundgesetz Deutschlands oder regulatorischen Vorgaben genügen, (sogenannte Constitutional AI): KI-Modelle lehren KI-Modelle ethische Regeln.

Was kann getan werden?

- Regulatorisch sollte es verboten werden, für Nutzer schädliche Modifikationen der verwendeten Basismodelle vorzunehmen.
- Die Entwicklung von KI ist zu wichtig, um sie alleine IT-Experten zu überlassen. Die Entwickler-Teams sollten daher eine gewisse Diversität aufweisen, die repräsentativ für das Unternehmen und/oder die Gesellschaft ist. Frauen sind leider derzeit in der KI-Entwicklung stark unterrepräsentiert.
- Wie die aktuelle Wirtschaftsnobelpreisträgerin, Claudia Goldin, herausgearbeitet hat, werden Frauen in den westlichen Industrieländern durch gesellschaftliche Strukturen systematisch daran gehindert, ihr Arbeitsangebot auszuweiten. Dies gilt auch für Deutschland. Die Diskriminierung von Frauen im Alltag ist wahrscheinlich bedeutsamer als die Diskriminierung durch KI.
- Modellentwicklungen, die explizit auf eine breitere Datenbasis zurückgreifen, die marginalisierte Gruppen stärker repräsentieren, sind zu begrüßen.
- Entwicklungen, die Transparenzdefizite von KI-Modellen aufarbeiten, sind voranzutreiben.
- Für sehr sensible gesellschaftliche Bereiche (beispielsweise Wohnen, Beschäftigung, Medizin) kann über die Pflicht zur Verwendung von Modellen nachgedacht werden, die Diskriminierungspotenziale minimieren.

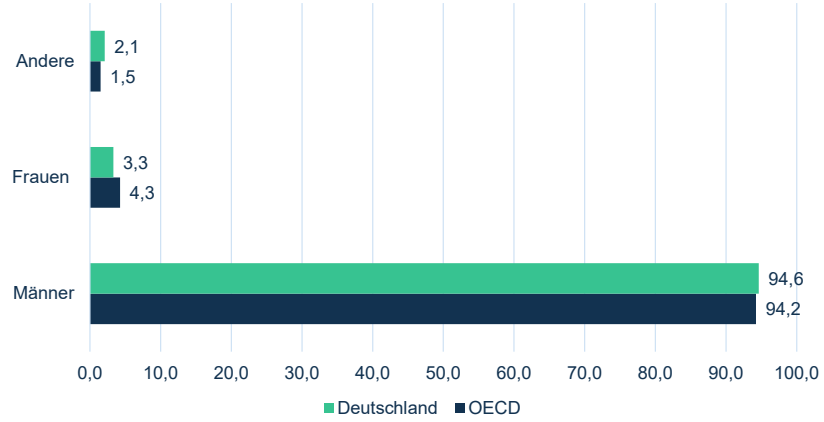
Diskriminierung von Frauen durch Steuerrecht stärker als durch KI

Training von GenAI-Modellen



Quelle: LBBW Research

Anteile von Geschlechtern, KI-Experten, in %, 2022



Quelle: OECD, LBBW Research

Fazit

Biases von KI-Modellen sind im Prinzip besser zu steuern als die Vorurteile von Menschen: Es sind die Menschen, die bei allen Schritten des Modellbaus ihre Vorurteile (unbewusst) in das Modell einpflegen.

Traditionelle KI-Modelle sind besser zu steuern als GenAI-Modelle. Biases in GenAI-Modellen sind nur schwer zu beheben – in erster Linie, weil die dahinter laufenden neuronalen Netze weitestgehend eine Black Box darstellen. Es wird viel Arbeit brauchen, die Diskriminierungspotenziale von GenAI-Modellen zu senken.

In Zukunft müssen GenAI-Modelle gebaut werden, die eine höhere Diversität in den Trainingsdaten aufweisen als bislang. Je diverser ein Unternehmen ist, desto mehr Bedeutung wird dies u.U. bekommen.

Disclaimer

Diese Publikation richtet sich ausschließlich an Empfänger in der EU, Schweiz und Liechtenstein.

Diese Publikation wird von der LBBW nicht an Personen in den USA vertrieben und die LBBW beabsichtigt nicht, Personen in den USA anzusprechen.

Aufsichtsbehörden der LBBW: Europäische Zentralbank (EZB), Sonnemannstraße 22, 60314 Frankfurt am Main und Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin), Graurheindorfer Str. 108, 53117 Bonn / Marie-Curie-Str. 24-28, 60439 Frankfurt.

Diese Publikation beruht auf von uns nicht überprüfbaren, allgemein zugänglichen Quellen, die wir für zuverlässig halten, für deren Richtigkeit und Vollständigkeit wir jedoch keine Gewähr übernehmen können. Sie gibt unsere unverbindliche Auffassung über den Markt und die Produkte zum Zeitpunkt des Redaktionsschlusses wieder, ungeachtet etwaiger Eigenbestände in diesen Produkten. Diese Publikation ersetzt nicht die persönliche Beratung. Sie dient nur zu Informationszwecken und gilt nicht als Angebot oder Aufforderung zum Kauf oder Verkauf. Für weitere zeitnähere Informationen über konkrete Anlagemöglichkeiten und zum Zwecke einer individuellen Anlageberatung wenden Sie sich bitte an Ihren Anlageberater.

Wir behalten uns vor, unsere hier geäußerte Meinung jederzeit und ohne Vorankündigung zu ändern. Wir behalten uns des Weiteren vor, ohne weitere Vorankündigung Aktualisierungen dieser Information nicht vorzunehmen oder völlig einzustellen.

Die in dieser Ausarbeitung abgebildeten oder beschriebenen früheren Wertentwicklungen, Simulationen oder Prognosen stellen keinen verlässlichen Indikator für die künftige Wertentwicklung dar.

Die Entgegennahme von Research Dienstleistungen durch ein Wertpapierdienstleistungsunternehmen kann aufsichtsrechtlich als Zuwendung qualifiziert werden. In diesen Fällen geht die LBBW davon aus, dass die Zuwendung dazu bestimmt ist, die Qualität der jeweiligen Dienstleistung für den Kunden des Zuwendungsempfängers zu verbessern.