Deepfakes: Eine Einordnung

Einleitung

Mit den fortschreitenden Entwicklungen im Bereich der künstlichen Intelligenz (KI) gewinnt die Verbreitung von Deepfakes an Popularität. Deepfakes sind digitale Medieninhalte, die authentisch wirken, es aber nicht sind. Sie sind keine eigenständige Technologie, sondern nutzen bestehende Technologien und Methoden aus dem Bereich der künstlichen Intelligenz. Ihre Nutzung erstreckt sich über verschiedene Bereiche, von der Unterhaltungsindustrie bis hin zu politischen und gesellschaftlichen Kampagnen. Ihre Auswirkungen sind tiefgreifend und vielseitig.

Die nachfolgende Einordnung des BVDW beleuchtet die technologische Entwicklung von Deepfakes, ihre vielfältigen Anwendungen und die damit verbundenen gesellschaftlichen und ethischen Herausforderungen, um ein umfassendes Verständnis zu fördern und Ideen zur Prävention und Bekämpfung zu entwickeln.

Technische Definition von Deepfakes

Deepfakes sind hochentwickelte digitale Medieninhalte, die mithilfe von KI und maschinellem Lernen (ML) manipuliert werden. Dazu zählen Video-, Audio- und Bildinhalte, die täuschend echt erscheinen, obwohl sie vollständig oder teilweise gefälscht sind. Der Begriff „Deepfake“ leitet sich von „deep learning“, einer fortgeschrittenen Methode des maschinellen Lernens, und „fake“ (englisch für Fälschung) ab. Diese Technologie nutzt komplexe neuronale Netzwerke, um visuelle und auditive Daten zu analysieren und zu rekonstruieren, wodurch realistische Fälschungen erzeugt werden können.

Gesellschaftliche Definition von Deepfakes

Deepfakes können in der modernen digitalen Gesellschaft das Vertrauen in mediale Inhalte fundamental erschüttern und weitreichende gesellschaftliche, politische und wirtschaftliche Auswirkungen haben.

Deshalb sind unter Deepfakes solche Veränderungen von medialen Inhalten zu verstehen, die grundsätzlich mit einer manipulativen Absicht erstellt werden. Ein Deepfake soll nicht von einem echten Inhalt unterscheitbar sein. Deshalb wird er auch in der Art der Verbreitung, dem Inhalt und dem Erscheinungsbild so gestaltet, dass er glaubwürdig ist.

Dies unterscheidet Deepfakes von anderen abgeänderten digitalen Medieninhalten, bspw. für künstlerische oder satirische Zwecke. Hier wird entweder durch die Form, den Inhalt oder die Platzierung erkenntlich, dass es sich um einen bearbeiteten digitalen Medieninhalt handelt.

Status Quo und Debatte

Mit der zunehmenden Verbreitung und des einfacheren Zugangs zu KI-Tools, die die Bearbeitung von digitalen Medieninhalten ermöglichen, wächst die Häufigkeit von Vorfällen, bei denen Deepfakes zum Einsatz kommen.

Deepfakes stellen in zahlreichen persönlichen und gesellschaftlichen Dimensionen eine erhebliche Bedrohung dar. Dazu zählen unter anderem Identitätsdiebstahl, Rufschädigung und die Verbreitung von Desinformationen. Studien zeigen, dass Deepfake-Angriffe um bis zu 3.000% zugenommen haben, insbesondere in Bereichen wie Kryptowährung und Online-Banking. Dies unterstreicht die Notwendigkeit von robusten Abwehrmaßnahmen[1] und verdeutlicht die dringende Notwendigkeit, die Entwicklung der Deepfake-Technologie sowohl zu verstehen, als auch einen rechtssicheren Rahmen zu schaffen.

Als erstes ist festzuhalten, dass legitime und manipulative digitale Inhalte technisch und in der Erstellung nicht zu unterscheiden sind. Für die gesellschaftliche Debatte braucht es daher unbedingt eine andere Begrifflichkeit für die legitime Bearbeitung von digitalen Medieninhalten. Dies ist Voraussetzung, um die positive Nutzung klar und präzise abgrenzen zu können und den Begriff Deepfakes lediglich für digitale Inhalte zu benutzen, die zu manipulativen Zwecken verändert werden.

1 Quelle: https://onfido.com/landing/identity-fraud-report/?ref=hackernoon.com

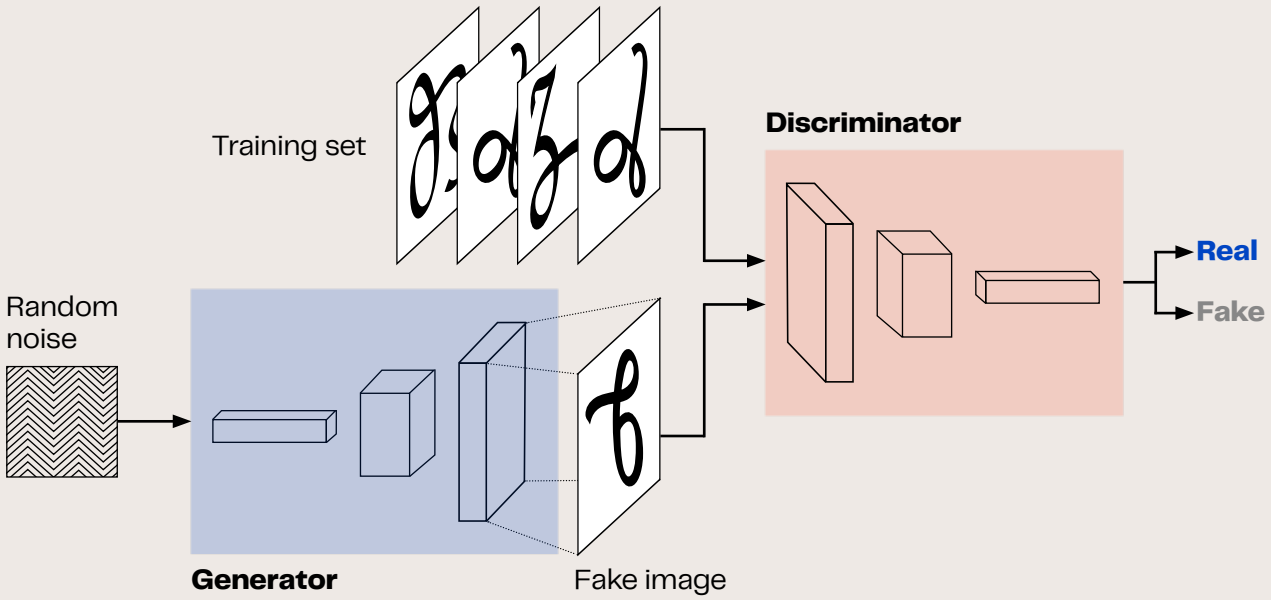
Hintergrund und Kontext

Entwicklung der Technologie

Die Entwicklung der Deepfake-Technologie begann mit den bedeutenden Fortschritten in der künstlichen Intelligenz und im maschinellen Lernen. Ein entscheidender Durchbruch war die Einführung der Generative Adversarial Networks (GANs) durch Ian Goodfellow im Jahr 2014. [2]

Ein GAN ist ein maschinelles Lernmodell, das aus zwei neuronalen Netzwerken besteht, die gegeneinander trainiert werden. Diese beiden Netzwerke werden als Generator und Diskriminator bezeichnet.

Generative Adversarial Networks (GANs)



Das Ziel des Generators ist es, neue Daten zu erzeugen, die so realistisch wie möglich aussehen. Der Generator nimmt einen zufälligen Eingangsvektor (oft Rauschen genannt) und wandelt diesen in eine synthetische, realistisch aussehende Ausgabe um.

Der Diskriminator erhält sowohl echte Daten (aus dem Trainingssatz) als auch die vom Generator erzeugten Daten und versucht zu unterscheiden, ob die Daten real oder künstlich sind. Sein Ziel ist es, echte Daten von gefälschten zu unterscheiden. Während des Trainingsprozesses spielen der Generator und der Diskriminator ein Minimax-Spiel gegeneinander. Der Generator versucht, den Diskriminator zu täuschen, indem er immer realistischere Daten erzeugt, während der Diskriminator versucht, besser im Erkennen von gefälschten Daten zu werden.

Der Generator wird somit trainiert, um den Verlust des Diskriminators zu maximieren und somit den Diskriminator zu täuschen. Der Diskriminator wird gleichzeitig trainiert, um den Verlust zu minimieren, das heißt echt und gefälscht korrekt zu klassifizieren. Das Ziel dieses Wettbewerbs ist es, dass der Generator schließlich so gute synthetische Daten erzeugt, dass der Diskriminator nicht mehr zuverlässig zwischen echten und gefälschten Daten unterscheiden kann. Wenn dies erreicht ist, hat das GAN erfolgreich gelernt, realistische Daten zu erzeugen. [3]

2 Quelle: https://the-decoder.de/geschichte-der-deepfakes-so-rasant-geht-es-mit-ki-fakes-voran/
3 Quelle: https://www.datacamp.com/tutorial/generative-adversarial-networks

Arten von Deepfakes

Es gibt verschiedene Arten von Deepfakes, die jeweils auf unterschiedliche Medien und Techniken abzielen. Übergeordnet sind insbesondere Video –, Audio – und Bild-Deepfakes verbreitet.

Arten von Deepfakes

Video-Deepfakes:

Diese beinhalten die Manipulation von Videoinhalten, bei denen Gesichter oder Körperbewegungen so verändert werden, dass Personen scheinbar Dinge sagen oder tun, die sie in Wirklichkeit nie gesagt oder getan haben.

Beispielsweise führte im Juni 2022 Berlins Bürgermeisterin Franziska Giffey ein Videotelefonat mit einer Person, die vorgab, Kiews Bürgermeister Vitali Klitschko zu sein. Das Gespräch verlief zunächst normal, bis der vermeintliche Klitschko ungewöhnliche Fragen stellte, was Giffeys Verdacht weckte. Nach Abbruch des Gesprächs bestätigte der ukrainische Botschafter, dass es sich um einen Betrug handelte, vermutlich mit Hilfe eines Deepfake. .[4]

Audio-Deepfakes:

Hierbei handelt es sich um gefälschte Audiodateien, bei denen Stimmen nachgeahmt werden, um falsche Aussagen zu erzeugen. Ein Vorfall hierfür fand während der letzten Wahlkampagne in Polen statt. Die politische Partei des Ministerpräsidenten Donald Tusk hat einen Deepfake-Audio-Clip seines Gegners in den sozialen Medien veröffentlicht. Diese Technologie wurde genutzt, um den politischen Gegner zu diskreditieren und Wählende zu beeinflussen.[5]

Bild-Deepfakes:

Diese beinhalten die Manipulation von Bildern, um Personen oder Szenen darzustellen, die so nie existierten. Beispielsweise kursierten im Januar 2024 eine Welle von gefälschten pornografischen Bildern im Netz, die angeblich Taylor Swift zeigen, welche auf Plattformen wie Twitter und Telegram verbreitet wurden. [6]

Herausforderungen und Risiken

Die schnelle Entwicklung der technischen Grundlage für Deepfakes bringt zahlreiche Herausforderungen und Risiken mit sich, die weit über die offensichtlichen Bedrohungen hinausgehen. Es ist relevant, die vielfältigen gesellschaftlichen und wirtschaftlichen Auswirkungen zu identifizieren und sich mit den ethischen und rechtlichen Aspekten auseinander zu setzen, die durch die Nutzung dieser Technologie entstehen. Eine erste übergreifende Herausforderung ist der rasante Fortschritt der Technologien zur Erstellung von Deepfakes. Dies führt zu einer Art Wettrüsten zwischen der erstellenden Software und der Analyse- und Erkennungssoftware von Deepfakes.

4 Quelle: https://www.dw.com/en/vitali-klitschko-fake-tricks-berlin-mayor-in-video-call/a-62257289
5 Quelle: https://www.politico.eu/article/deepfakes-distrust-disinformation-welcome-ai-election-2024/
6 Quelle: https://www.ijpr.org/media-society/2024-01-27/deepfakes-exploiting-taylor-swift-images-exemplify-a-scourge-with-little-oversight

Erkennungsmechanismen

Die Erkennung von Deepfakes ist eine komplexe Herausforderung, die eine Kombination aus technologischen, algorithmischen und menschlichen Ansätzen erfordert. Aktuelle Methoden zur Erkennung von Deepfakes umfassen eine Vielzahl von algorithmischen Ansätzen, die auf spezifischen Mustern in der Bild- oder Tonverarbeitung basieren. Ein gängiger Ansatz ist die Analyse von Unregelmäßigkeiten in der Beleuchtung und Schatten der Lichtverhältnisse in gefälschten Videos, die für das menschliche Auge schwer zu erkennen, aber für Algorithmen offensichtlich sind. [7]

Fortgeschrittene KI-Modelle wie Convolutional Neural Networks (CNNs) werden trainiert, um subtile Anomalien in der Gesichtsbewegung oder Sprachmodulation zu erkennen, die auf eine Fälschung hinweisen könnten. Neuere Techniken verwenden auch die Analyse von Blinzelmustern, da diese oft in Deepfakes fehlen oder unnatürlich erscheinen. Zudem werden fortschrittliche Forensik-Tools entwickelt, die auf die Metadaten von Mediendateien zugreifen, um Hinweise auf Manipulationen zu finden.

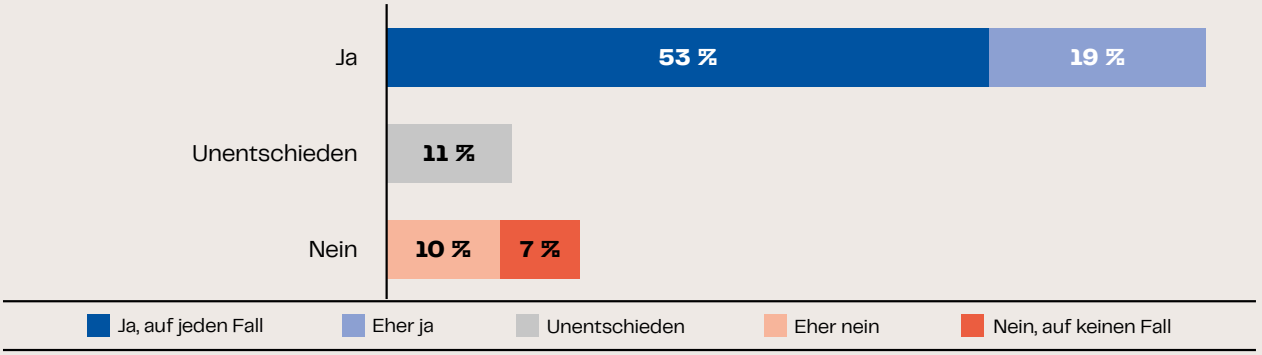
Trotz der Fortschritte in der Erkennungstechnologie bleibt die Bekämpfung von Deepfakes eine technische Herausforderung. Die kontinuierliche Verbesserung der Fälschungstechniken bedeutet, dass Erkennungsalgorithmen ständig aktualisiert werden müssen. Darüber hinaus erfordert die Analyse von Mediendateien erhebliche Rechenressourcen und kann zeitaufwendig sein.

Gesellschaftliche Auswirkungen

Technologieanwendungen, die es ermöglichen, realistische, aber gefälschte Video- und Audioinhalte zu erstellen, haben weitreichende gesellschaftliche Auswirkungen. Deepfakes können zur Verbreitung falscher Informationen genutzt werden, was das Vertrauen in Medien und öffentliche Institutionen untergräbt. In politischen Kontexten können Deepfakes eingesetzt werden, um Wahlentscheidungen zu beeinflussen und den Wahlkampf zu manipulieren, politische Gegner zu diskreditieren oder gesellschaftliche Unruhen zu fördern. Zudem stellen sie eine erhebliche Bedrohung für die Privatsphäre dar. Deepfakes können Menschen in kompromittierenden Situationen darstellen, die nie stattgefunden haben. Dies kann von Rufschädigung bis zu sozialer Isolation führen. Prominente und Bürger*innen sind gleichermaßen gefährdet, Opfer solcher Techniken zu werden.

In einer repräsentativen Civey-Befragung im Auftrag des BVDW sagen 72 Prozent der 2.500 Befragten, dass das Vertrauen in digitale Medien durch die Verbreitung von Deepfakes definitiv oder eher sinkt. Dies hat signifikante Auswirkungen auf die digitale Wirtschaft. Ein Rückgang des Vertrauens in digitale Medien kann zu einer verminderten Nutzung von Digitalen Diensten, Angeboten und Plattformen führen. Unweigerlich wird dies negative Auswirkungen auf digitale Geschäftsmodelle haben. Unternehmen, die auf digitale Medien angewiesen sind, müssen daher verstärkt in Maßnahmen zur Sicherung der Authentizität und Transparenz ihrer Inhalte investieren.

Beeinflussung des allgemeinen Vertrauens digitaler Medien durch die Verbreitung von Deepfakes



Frage 1: Beeinflusst die Verbreitung von „Deepfakes" (täuschend echt wirkende, manipulierte Bild-, Audio- oder auch Videoaufnahmen) Ihr allgemeines Vertrauen in digitale Medien? | Stat. Fehler Gesamtergebnis: 3,4 % | Stichprobengröße: 2.507 | Befragungszeitraum: 12.06.2024 - 13.06.2024

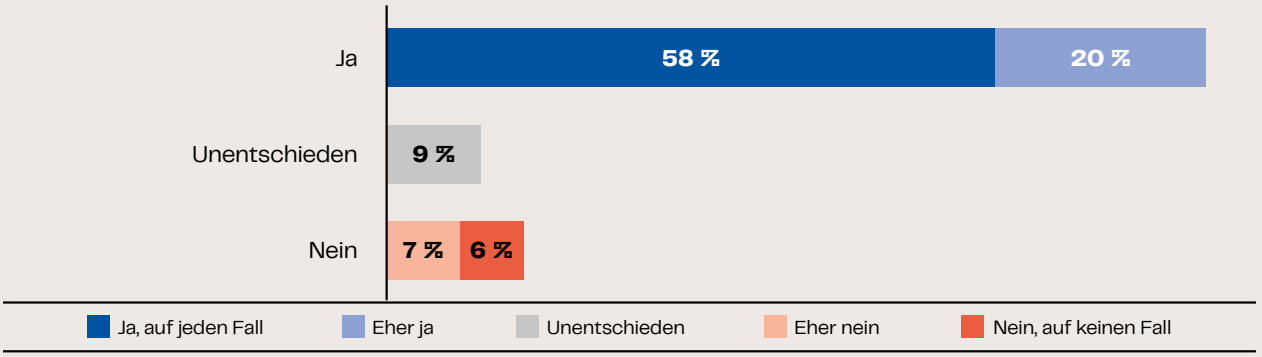
Civey

7 Quelle: https://www.theguardian.com/us-news/article/2024/jun/07/how-to-spot-a-deepfake?utm_source=%2Fmeta%2Fnewsflash%2Fvor9&utm_medium=newsletter&utm_campaign=nl3134&utm_term=a4755db027dbefd6fb1a09566a9e8e6b

Zudem geben 78 % aller Teilnehmer*innen an, dass auch die Glaubwürdigkeit von legitimen Nachrichten und Informationen durch die Verbreitung von Deepfakes beeinträchtigt wird. Dies deutet auf eine umfassende Herausforderung für die gesamte Informationsökonomie hin. Wenn legitime Nachrichten und Informationen weniger glaubwürdig erscheinen, könnte dies das Vertrauen der Öffentlichkeit in Medien und Informationsquellen insgesamt untergraben. Dieses Misstrauen führt dazu, dass Menschen zunehmend alternative Informationsquellen suchen, die möglicherweise weniger vertrauenswürdig oder irreführend sind. Dies führt zu einer verstärkten Verbreitung von Fehlinformationen und könnte die Fähigkeit der Bürger*innen, fundierte Entscheidungen zu treffen, beeinträchtigen.

Für Unternehmen bedeutet dies, dass sie möglicherweise mehr Ressourcen in die Verifizierung und Authentifizierung von Informationen investieren müssen, um ihre Reputation zu schützen und das Vertrauen von Kund*innen zu bewahren. Langfristig könnte dies einen Wandel hin zu strengeren Standards und höheren Anforderungen an die digitale Kommunikation und Informationsverbreitung bedeuten.

Beeinflussung der Glaubwürdigkeit von Nachrichten und Informationen durch die Verbreitung von Deepfakes



Frage 2: Beeinflusst die Verbreitung von „Deepfakes" (täuschend echt wirkende, manipulierte Bild-, Audio- oder auch Videoaufnahmen) Ihrer Meinung nach die Glaubwürdigkeit von Nachrichten und Informationen?
Stat. Fehler Gesamtergebnis: 3,5 % | Stichprobengröße: 2.506 | Befragungszeitraum: 12.06.2024 – 13.06.2024

Civey

Wirtschaftliche Herausforderungen

Deepfakes stellen auch erhebliche wirtschaftliche Herausforderungen dar. Diese reichen von finanziellem Betrug bis hin zu den Kosten für den Schutz vor Deepfake-Angriffen. Unternehmen müssen beträchtliche Ressourcen aufwenden, um sich gegen solche Bedrohungen zu wappnen. Gleichzeitig können Deepfakes das Vertrauen der Nutzer*innen in digitale Dienstleistungen und Finanztransaktionen untergraben.

Ein weiterer wirtschaftlicher Aspekt ist der Verlust an Markenvertrauen. Wenn gefälschte Videos oder Bilder von Unternehmensleitern oder Prominenten in Verbindung mit einer Marke verbreitet werden, kann dies zu erheblichen Reputationsschäden führen und das Vertrauen der Verbraucher*innen in die Marke beeinträchtigen. Dies kann letztlich zu Umsatzverlusten und einem Rückgang der Kundenbindung führen.

Verantwortlichkeit und ethische Überlegungen

Die Erstellung und Verbreitung von Deepfakes werfen daneben auch erhebliche ethische Fragen auf. Entwickler*innen und Plattformen, die diese Technologien bereitstellen, tragen eine große Verantwortung. Es ist notwendig, ethische Richtlinien zu entwickeln, um- und durchzusetzen, die den Missbrauch von Deepfakes verhindern. Dazu gehört auch, dass Unternehmen und Entwickler*innen transparent arbeiten und Maßnahmen ergreifen, um die Verbreitung schädlicher Inhalte zu verhindern.

Fazit und Ausblick

Die technologische Entwicklung und die zunehmende Verbreitung von Deepfakes stellen Gesellschaft und Wirtschaft vor erhebliche Herausforderungen. Es ist zu erwarten, dass die technologische Entwicklung von Deepfakes weiterhin Fortschritte macht und Fälschungen immer schwerer zu erkennen sind.

Gleichzeitig werden neue Erkennungstechnologien und –methoden entwickelt, um Schritt zu halten. Die Forschung wird sich auf die Verbesserung der Erkennungsgenauigkeit und die Entwicklung robuster Abwehrmechanismen konzentrieren müssen.

Um den Herausforderungen und Risiken durch Deepfakes zu begegnen, ist es essenziell, dass verschiedene Akteure aus Wirtschaft, Politik und Gesellschaft gemeinsam und koordiniert Maßnahmen ergreifen.

Es bedarf einer ausgewogenen Herangehensweise, die sowohl die gesellschaftlichen und wirtschaftlichen Risiken im Blick hat, als auch die Unterstützung technologischer Innovationen sicherstellt. Daher müssen die schlechten Anwendungsfälle – wie in diesem Papier in groben Zügen skizziert – klar benannt und definiert werden. Um den gesellschaftlichen Auswirkungen entgegenzutreten, bedarf es umfassender Aufklärung, ohne in Alarmismus zu verfallen. Die Steigerung der Medienkompetenz in allen Altersgruppen ist essenziell, um gegen die Verbreitung von Deepfakes vorzugehen. Dabei kann können Plattformen mitwirken, auf denen solche Inhalte geteilt werden und Nutzer*innen dazu aufrufen, zweimal zu prüfen, ob ein Inhalt glaubwürdig ist und daher weiterverbreitet werden sollte.

Autor*innen

Kai Ebert
Director Growth, SYZYGY GROUP

Katharina Jäger
Head of Innovation & Technology, BVDW

Jens-Christian Jensen
Chief Strategy Officer, Plan.Net at Serviceplan Group

Janek Kuberzig
Public Affairs Manager, BVDW

Bundesverband Digitale Wirtschaft (BVDW) e.V.

Der Bundesverband Digitale Wirtschaft (BVDW) e.V. ist die Interessenvertretung für Unternehmen, die digitale Geschäftsmodelle betreiben oder deren Wertschöpfung auf dem Einsatz digitaler Technologien beruht. Als Impulsgeber, Wegweiser und Beschleuniger digitaler Geschäftsmodelle vertritt der BVDW die Interessen der Digitalen Wirtschaft gegenüber Politik und Gesellschaft und setzt sich für die Schaffung von Markttransparenz und innovationsfreundlichen Rahmenbedingungen ein. Sein Netzwerk von Experten liefert mit Zahlen, Daten und Fakten Orientierung zu einem zentralen Zukunftsfeld. Neben der DMEXCO und dem Deutschen Digital Award richtet der BVDW eine Vielzahl von Fachveranstaltungen aus. Mit Mitgliedern aus verschiedensten Branchen ist der BVDW die Stimme der Digitalen Wirtschaft.

Ressort Künstliche Intelligenz

Die gewinnbringende und verantwortungsvolle Nutzung von künstlicher Intelligenz (KI) in der deutschen digitalen Wirtschaft steht im Fokus der Ressortarbeit. Ziel ist es, Fragen rund um die Veränderungen der Wertschöpfungskette der digitalen Wirtschaft zu beantworten und Lösungsansätze für die ethischen, sozialen und rechtlichen Herausforderungen durch KI zu bieten, um eine nachhaltige und positive Auswirkung auf die Gesellschaft, Wirtschaft und Umwelt sicherzustellen.

Kontakt

Katharina Jäger, Head of Innovation & Technology, jaeger@bvdw.org

Bundesverband Digitale Wirtschaft (BVDW) e.V.
Schumannstraße 2, 10117 Berlin
www.bvdw.org