digiLabs: Digitalisierung und Künstliche Intelligenz gegen Antisemitismus

Chancen · Methoden · Werkzeuge · neue Bedrohungen

iibsa
Research & Education on Antisemitism

digiLabs
Digitalisierung und Künstliche Intelligenz gegen Antisemitismus.

Herausgegeben von
Internationales Institut für Bildung, Sozial- und Antisemitismusforschung e.V.
Registriert beim Amtsgericht Berlin-Charlottenburg. Vertreten durch die Vorstandsvorsitzende Dr. Kim Robin Stoller.
Pariser Platz 6a, 10117 Berlin · Telefon: 030 55 214 934 · mail@iibsa.org · www.iibsa.org

Redaktion und V.i.S.d.P.
Dr. Kim Robin Stoller, Internationales Institut für Bildung, Sozial- und Antisemitismusforschung e.V.

Haftungsausschluss
Die Informationen in dieser Publikation wurden nach bestem Wissen und Gewissen formuliert. Für die Vollständigkeit und Aktualität der Informationen übernehmen die Herausgeber keine Gewähr. Die Publikation enthält Hinweise auf Webseiten und Werkzeuge Dritter, auf deren Inhalt wir keinen Einfluss haben. Für diese Inhalte übernehmen wir daher keine Gewähr. Für die Inhalte der angegebenen oder verlinkten Seiten ist stets der/die jeweilige Anbieter:in oder Betreiber:in verantwortlich. Die genannten Werkzeuge dienen ausschließlich der Veranschaulichung; ihre Nennung stellt keine Empfehlung dar. Für inhaltliche Aussagen tragen die Autor_innen die Verantwortung.

Wichtiger Hinweis: Diese Handreichung ersetzt keine rechtliche Beratung. Vor dem Einsatz der beschriebenen Methoden – insbesondere bei der Erhebung, Verarbeitung und Veröffentlichung personenbezogener Daten – ist im Zweifel eine fachkundige rechtliche Einschätzung einzuholen.

Inhalt

Über diese Publikation

Antisemitismus verändert sich – und mit ihm die Werkzeuge, die zu seiner Analyse und Bekämpfung zur Verfügung stehen. Digitalisierung und Künstliche Intelligenz (KI) eröffnen neue Möglichkeiten, antisemitische Inhalte zu erkennen, ihre Verbreitung zu verstehen und Aufklärungsarbeit wirksamer zu gestalten. Zugleich werden dieselben Technologien von antisemitischen Akteur:innen genutzt, um Angriffe durchzuführen und Hass zu verbreiten. Wer im Feld der Antisemitismusbekämpfung arbeitet, steht damit vor einer doppelten Aufgabe: die neuen Technologien zu verstehen und sie verantwortungsvoll einzusetzen.

Diese Handreichung bündelt praxisnahes Wissen aus dem Projekt DigiLabs des IIBSA, das zivilgesellschaftliche Akteur:innen, jüdische Organisationen und Expert:innen aus Wissenschaft, Recht und Cybersicherheit zusammengebracht hat. Sie richtet sich ausdrücklich auch an Einsteiger:innen ohne technische Vorkenntnisse.

Das Projekt DigiLabs

Diese Handreichung entstand im Rahmen des Projekts „DigiLabs – Digitalisierung und Künstliche Intelligenz gegen Antisemitismus“ des Internationalen Instituts für Bildung, Sozial- und Antisemitismusforschung e.V. (IIBSA). Das Projekt wurde gefördert vom Beauftragten der Bundesregierung für jüdisches Leben in Deutschland und den Kampf gegen Antisemitismus. Über eine Laufzeit von mehr als zwei Jahren (2024–2026) verfolgte das Projekt ein Ziel: die Lücke zwischen traditionellen Methoden der Analyse und Bekämpfung von Antisemitismus und den Anforderungen einer von Digitalisierung und KI geprägten Welt zu schließen und zivilgesellschaftliche und wissenschaftliche Akteur:innen zu befähigen, digitale Werkzeuge wirksam und verantwortungsvoll einzusetzen.

Kern des Projekts waren zehn Workshops in fünf Themenfeldern sowie ein Austausch-Forum. Die Angebote richteten sich ausdrücklich auch an Einsteiger:innen ohne technische Vorkenntnisse. Sie brachten zivilgesellschaftliche Akteur:innen, jüdische Organisationen und Expert:innen aus Wissenschaft, Recht und Cybersicherheit zusammen, um die Herausforderungen zu diskutieren und gemeinsam innovative Lösungen zu erproben.

Das Projekt auf einen Blick
Laufzeit: März 2024 bis August 2026
Maßnahmen: 10 Workshops in 5 Themenfeldern + DigiLabs-Austausch-Forum
Zielgruppen: zivilgesellschaftliche Akteur:innen und Multiplikator:innen gegen Antisemitismus, jüdische Organisationen und Gemeinden, Akteur:innen aus KI, Informatik und Digitalisierung
Förderung: Beauftragter der Bundesregierung für jüdisches Leben in Deutschland und den Kampf gegen Antisemitismus

:: Die Workshops im Überblick

Das Themenspektrum reichte von Einstiegsworkshops in KI und Large Language Models über die quantitative und qualitative Analyse von Antisemitismus, die Nutzung von KI und Digitalisierung für Recherche, Social Media und Kampagnen bis hin zu rechtlichen Fragen. Darüber hinaus

beschäftigten sich die Workshops mit Themen wie der Nutzung von KI durch antisemitische Akteur:innen und zukünftigen Herausforderungen. Zudem wurde der Einsatz von KI in der Berichterstellung erläutert. Durchgeführt wurden sie von ausgewiesenen Expert:innen aus Wissenschaft, Recht und Cybersicherheit.

:: Das DigiLabs-Austausch-Forum

Im Rahmen des Projekts wurde das DigiLabs-Austausch-Forum ins Leben gerufen als Ort, der Antisemitismus-Expertise und digitale Expertise zusammendenkt. Im Rahmen des Forums werden Bedarfe und Potenziale der KI-Nutzung in der Analyse und Bekämpfung von Antisemitismus erläutert.

1 · Antisemitismus im digitalen Zeitalter

Im digitalen Zeitalter verbreiten sich Informationen schneller als je zuvor. Soziale Netzwerke, Online-Foren, Messenger-Dienste und Videoplattformen ermöglichen Austausch über geografische Grenzen hinweg – werden aber auch genutzt, um antisemitische Ressentiments, Hassbotschaften und Verschwörungserzählungen zu verbreiten und Einflusskampagnen durchzuführen. Erhebungen zum digitalen Raum gehen davon aus, dass soziale Medien kombiniert mit Ereignissen und Kampagnen als Katalysatoren von antisemitischen Inhalten fungieren. Das Hamas-Massaker am 7. Oktober 2023 und darauffolgende kriegerische Auseinandersetzungen bilden dabei den Rahmen (Frame): Antisemitismus und Israelhass haben in sozialen Medien, an Hochschulen und im Kontext anti-israelischer Demonstrationen massiv zugenommen.

:: KI als Gefahr – und als Werkzeug

Antisemitische Akteur:innen nutzen KI, Digitalisierung und Automatisierung, um Hass zu verbreiten, die öffentliche Meinung zu beeinflussen und Fake News zu erstellen. Manipulierte und KI-generierte Bilder – sogenannte Deepfakes – finden seit dem 7. Oktober 2023 in beispiellosem Ausmaß Verwendung und verbreiten sich vor allem auf Plattformen wie X, TikTok, Instagram und Telegram. Die Qualität solcher Fälschungen ist mittlerweile so hoch, dass selbst Fachleute sie teilweise kaum noch als künstlich erkennen.

Teile der KI-Technologien lassen sich jedoch auch für die Gegenwehr nutzen. Moderne KI-Systeme können große Mengen an Text-, Bild- und Videodaten analysieren und Muster erkennen, die auf antisemitische Aussagen, Symbole oder Narrative hinweisen – schneller als eine ausschließlich manuelle Prüfung. Die menschliche Bewertung bleibt unverzichtbar, da Sprache oft mehrdeutig ist und der Kontext über die Bedeutung entscheidet.

:: Aktuelle Entwicklung: KI selbst als Risikofaktor

Die Risiken durch KI werden inzwischen in Fachkreisen, Politik und Medien breit diskutiert. Sicherheitsbehörden, Forscher:innen und zivilgesellschaftliche Organisationen berichten, dass extremistische Gruppen KI-Werkzeuge – Chatbots, generative Bildmodelle und Deepfakes – gezielt einsetzen, um antisemitische Propaganda zu automatisieren, ihre Reichweite zu erhöhen und die Selbstradikalisierung Einzelner zu fördern. Dass selbst etablierte Systeme entgleisen können, zeigte der KI-Chatbot Grok (xAI), der im Sommer 2025 nach einem Update antisemitische Inhalte ausgab, sich zeitweise „MechaHitler“ nannte und den Holocaust relativierte. Das verdeutlicht: KI ist nicht per se ein Werkzeug der Aufklärung – ihr Einsatz erfordert Sorgfalt, Kontrolle und kritische Begleitung. Agentische KI, also Modelle, die lange Aufgabenketten ohne menschliches Zutun abarbeiten können, stellt eine zunehmende Gefahr dar. Cyberangriffe, Ausspähungen, Doxxing, Terroranleitungen und Kampagnen werden nun durch eine Vielzahl staatlicher und nicht-staatlicher antisemitischer Akteursgruppen möglich.

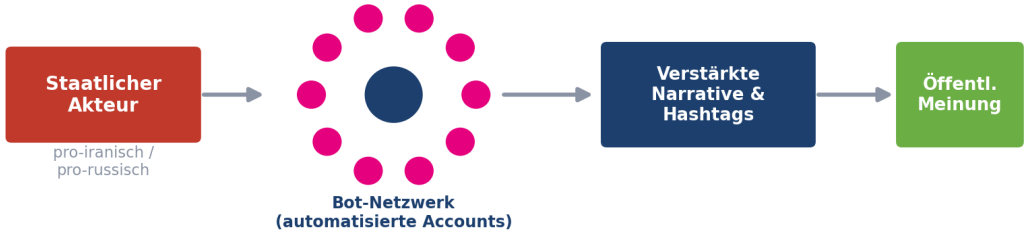
2 · Neue Bedrohungen: Einflussoperationen und Bot-Netzwerke

Dieselben Technologien, die der Analyse dienen, werden auch offensiv eingesetzt. Besonders folgenreich sind koordinierte Einflussoperationen, die KI und Automatisierung nutzen, um antisemitische und israelfeindliche Narrative massenhaft zu verbreiten und die öffentliche Meinung zu beeinflussen.

:: Wie Bot-Netzwerke wirken

Bei „koordiniertem unauthentischem Verhalten“ steuert ein zentraler Akteur eine Vielzahl automatisierter oder halbautomatischer Accounts. Diese kommentieren, posten und teilen abgestimmt Inhalte, um bestimmte Hashtags und Narrative künstlich zu verstärken, bis sie authentische Nutzer:innen und die öffentliche Debatte erreichen. KI senkt die Kosten dramatisch: Texte, Bilder und sogar Deepfake-Videos lassen sich automatisiert in großer Zahl und in vielen Sprachen erzeugen.

KI-gestützte Einflussoperationen: wie Bot-Netzwerke wirken



Was das für die Praxis bedeutet

Reichweite ist nicht gleich Resonanz. Auffällig gleichförmige Verbreitung, plötzliche Hashtag-Wellen und koordinierte Accounts können auf Einflussoperationen hindeuten – Manipulation ist als Möglichkeit stets einzukalkulieren.

:: Reichweite auf Knopfdruck: Collabs und Kampagnen-Accounts

Eine zentrale Beobachtung aus dem DigiLabs-Forum: Die Social-Media-Präsenz antisemitischer und israelfeindlicher Inhalte gewinnt enorm an Bedeutung – nicht zuletzt durch „Collabs“ (Kollaborationen) und eigens eingerichtete Kampagnen-Accounts. Über Collab-Funktionen erscheint ein einzelner Beitrag gleichzeitig auf mehreren Profilen und erreicht schlagartig deren gesamtes Publikum; neue oder bis dahin unbekannte Accounts können so „von null auf hundert“ sichtbar werden. In der im Forum vorgestellten Hochschulstudie entstanden bis zu 40 Prozent der Beiträge über solche Collab-Funktionen. Kampagnen-Accounts orchestrieren diese Dynamik: Sie posten koordiniert, verstärken sich gegenseitig und treiben Hashtags und Narrative gezielt nach oben. Für die Analyse heißt das: Nicht nur einzelne Inhalte zählen, sondern vor allem die Vernetzung und das gemeinsame Posten der Accounts.

:: Verstärkung durch Algorithmen: wenn Empörung belohnt wird

Die Reichweite problematischer Inhalte wird nicht nur künstlich erzeugt, sondern auch strukturell begünstigt. Die Empfehlungsalgorithmen großer Plattformen sind auf Verweildauer und Interaktion („Engagement") optimiert – und genau das belohnt zum Teil emotional aufgeladene, empörende und polarisierende Inhalte. Studien zeigen, dass engagement-basierte Ranking-Systeme feindselige, das „andere Lager" abwertende Beiträge bevorzugt ausspielen, obwohl Nutzer:innen solche Inhalte nicht einmal bevorzugen. Ein kontrolliertes Experiment während des US-Wahlkampfs 2024 lieferte sogar kausale Belege dafür, dass eine stärkere Ausspielung solcher Inhalte die affektive Polarisierung messbar erhöht.

Für die Verbreitung von Antisemitismus ist das doppelt problematisch. Antisemitische Hetze, Verschwörungserzählungen und drastische Israel-Dämonisierungen sind emotional zugespitzt und erzeugen viel Interaktion – und werden dadurch algorithmisch verstärkt, wenn durch die Plattformen nicht gezielt gegengesteuert wird. Differenzierte Einordnungen werden sonst teilweise strukturell benachteiligt. Hinzu kommen Echokammer-Effekte: Wer bestimmte Inhalte konsumiert, bekommt zunehmend ähnliche ausgespielt, was Radikalisierung begünstigen kann. Bot-Netzwerke, Collab-Kampagnen und algorithmische Verstärkung greifen so ineinander.

:: Pro-iranische, pro-russische und islamistische Operationen

Untersuchungen seit dem 7. Oktober 2023 dokumentieren staatlich gesteuerte Kampagnen. Im Krieg zwischen Israel und Iran (Juni 2025) flutete ein mutmaßlich der iranischen Revolutionsgarde (IRGC) nahestehendes Netzwerk die Plattform X mit antisemitischen Verschwörungserzählungen – etwa, eine „jüdische Lobby" ziehe die USA in den Krieg – und war Berichten zufolge zeitweise für bis zu 60 Prozent des Traffics auf zentralen Kriegs-Hashtags verantwortlich. Auffällig: Als während des Krieges das iranische Internet ausfiel, verstummten auch diese Bot-Netzwerke – ein starkes Indiz für die Steuerung.

Eine Untersuchung des Forschungsinstituts IIBSA mit dem ARD-Faktenfinder kam zu dem Schluss, dass pro-iranische Bot-Netzwerke massiv Social-Media-Accounts staatlicher Fernsehsender, wie der Tagesschau, mit antisemitischen und Pro-Regime-Kommentaren fluten und so die öffentliche Meinung manipulieren.

Auch pro-russische Operationen sind seit dem russischen Angriffskrieg gegen die Ukraine aktiv und verbinden sich teils mit anti-israelischen Aktivitäten. Nach dem 7. Oktober 2023 verstärkten staatliche Akteure aus Iran, Russland und weiteren Ländern pro-Hamas-Inhalte und antisemitische Narrative.

Allgemein war zu beobachten, dass KI-generierte Bilder und Videos – etwa fabrizierte Aufnahmen vermeintlicher Angriffe Israels auf palästinensische Zivilist:innen – gezielt eingestreut wurden. Islamistische Akteure verbreiteten auch Videos, die eine militärische Mobilisierung der Armeen muslimischer Länder gegen Israel behaupteten.

3 · Kontrollierbarkeit, offene Modelle und „harmful“-Risiken

Ob KI der demokratischen Zivilgesellschaft nützt oder schadet, hängt entscheidend davon ab, wie kontrollierbar die zugrunde liegenden Modelle sind. Hier zeichnet sich ein grundlegendes Spannungsfeld ab.

:: Proprietär oder offen?

Proprietäre Modelle laufen in der Cloud der Anbieter. Sie verfügen über Sicherheitsfilter, werden zentral aktualisiert und überwacht, und ihre Nutzungsregeln lassen sich – zumindest theoretisch – durchsetzen. Offene Modelle (open-weight) hingegen können heruntergeladen, lokal betrieben und frei weitertrainiert werden. Das ist aus Datenschutzsicht ein Vorteil – schafft aber ein weiteres Kontrollproblem.

Kontrollierbarkeit: proprietäre vs. offene Modelle

Proprietäre Modelle (Cloud)	Offene Modelle (lokal)
• Sicherheitsfilter aktiv	• Filter per Fine-Tuning /
• zentral kontrolliert	• „Abliteration" entfernbar
• Nutzungsregeln	• lokal gehostet
• durchsetzbar	• keine zentrale Kontrolle
• Updates & Monitoring	

:: Wenn Schutzmechanismen entfernt werden

Aktuelle Forschung zeigt, dass sich die Sicherheitsvorkehrungen offener Modelle mit überschaubarem Aufwand entfernen lassen. Eine als „Abliteration“ bekannte Technik kann die Schutzfilter außer Kraft setzen, ohne die Leistungsfähigkeit bedeutend zu mindern. Dabei werden gezielt interne Repräsentationen entfernt, die für Verweigerungen zuständig sind. Frei verfügbare Werkzeuge automatisieren dies inzwischen für Hunderte Modelle. So entstehen „unzensierte“ Modelle, die bereitwillig auch schädliche Inhalte erzeugen – und bereits für Cyberkriminalität genutzt werden.

:: Automatisierte Inhalte – auch lokal gehostet

Ein wachsendes Problem ist die vollautomatische Erzeugung von Inhalten. In Verbindung mit Automatisierung und autonomen Agenten lassen sich Texte, Bilder und Videos in großer Zahl und ohne menschliche Kontrolle produzieren. Werden dafür offene Modelle lokal – also auf eigenen Rechnern – betrieben, entfällt jede zentrale Moderation: Es gibt keine Plattform, die eingreifen könnte, und keinen Anbieter, der Nutzungsregeln durchsetzen könnte. So können antisemitische Propaganda, Fake News und Deepfakes automatisiert und weitgehend unentdeckt erzeugt und verbreitet werden.

4 · Cyberangriffe, KI, Doxxing und Social Engineering

:: KI-gestützte Cyberangriffe und Social Engineering als Sicherheitsrisiko

KI senkt nicht nur die Schwelle für Desinformation, sondern auch für Cyberangriffe. Mit generativer KI lassen sich täuschend echte Phishing-Nachrichten, gefälschte Stimmen (Deepfake-Voice) und Schadsoftware mit geringem Aufwand und ohne tiefe technische Kenntnisse erzeugen. Sicherheitsfachleute beobachten einen sprunghaften Anstieg KI-gestützter Phishing-Angriffe; entsprechende Werkzeuge sind bereits für geringe Beträge erhältlich. Staatliche Akteure – darunter Iran, Russland, China und Nordkorea – haben ihre Cyberangriffe deutlich ausgeweitet.

Für jüdische Gemeinden, israelbezogene Organisationen und Einzelpersonen ist das kein abstraktes Risiko. Das Bundesamt für Verfassungsschutz warnt seit Längerem vor laufenden iranischen Cyberangriffen, die sich gezielt gegen jüdische und pro-israelische Einrichtungen, Nahost-Expert:innen und die oppositionelle iranische Exil-Community in Deutschland richten. Typisch ist ein „Social Engineering"-Vorgehen: Über Wochen wird – etwa mit Interview- oder Konferenzanfragen – Vertrauen aufgebaut, um Betroffene zum Öffnen von Schadsoftware zu bewegen, mit der sich E-Mails, Dokumente, Kontakt- und Kalenderdaten ausleiten lassen. Der Verfassungsschutz geht von einer fortbestehend hohen abstrakten Gefährdung für Israel- und USA-nahe sowie jüdische Einrichtungen wie Schulen und Synagogen aus. 2025 wurde zudem in Dänemark ein mutmaßlicher Spion festgenommen, der jüdische Personen und Einrichtungen in Berlin im Auftrag eines iranischen Geheimdienstes ausgespäht haben soll. Ziel soll die gezielte Ermordung bestimmter Personen gewesen sein.

:: Doxxing: das gezielte Offenlegen persönlicher Daten

Eine weitere Gefahr für Einzelpersonen ist Doxxing – das Zusammentragen und Veröffentlichen privater Informationen wie Adresse, Arbeitsplatz oder familiärem Umfeld, um Betroffene einzuschüchtern, zu bedrohen oder Angriffe zu ermöglichen. KI und automatisierte OSINT-Verfahren beschleunigen das Zusammenführen verstreuter öffentlicher Daten zu einem detaillierten Personenprofil erheblich. Besonders betroffen sind jüdische Aktive, Journalist:innen und Engagierte gegen Antisemitismus. Das gezielte Verbreiten personenbezogener Daten zum Schaden einer Person kann je nach Fall auch strafrechtlich relevant sein. Auf die Individuen hat das massive psychische und sicherheitsrelevante Folgen.

:: Schutz für Organisationen, Gemeinden und Einzelpersonen

- **Cybervorsicht:** Vorsicht bei unerwarteten Anhängen und Links, Zwei-Faktor-Authentifizierung, aktuelle Software; ungewöhnliche Anfragen über einen zweiten Kanal verifizieren.
- **Digitale Spur überprüfen:** die eigene öffentliche Datenspur regelmäßig prüfen und minimieren („Self-Doxxing-Check"), sensible Daten verschlüsseln, sichere Kommunikationswege nutzen.
- **Strukturell:** Sensibilisierung im Team, klare Melde- und Notfallwege sowie die Zusammenarbeit mit spezialisierten Stellen.

Konsequenz

Technische Schutzmechanismen allein genügen nicht. Umso wichtiger sind Medien- und KI-Kompetenz, wachsame Zivilgesellschaft, verbindliche Plattform- und Rechtsdurchsetzung sowie Fachwissen darüber, wie missbräuchlich erzeugte Inhalte, Kampagnen und Angriffe erkannt und eingeordnet werden können.

5 · Antisemitismus erkennen

Jede Analyse – ob manuell oder KI-gestützt – setzt voraus, dass man weiß, wonach man sucht. Antisemitismus lässt sich entlang verschiedener Dimensionen verstehen und analysieren: Kognition, Stereotype, Emotion und Sprache. Der Sprache kommt dabei eine herausragende Rolle zu, denn sie tradiert antisemitisches Denken und Fühlen. Antisemitismus tritt sowohl intentional auf (absichtliche Diffamierung) als auch nicht-intentional (unbewusste Stereotypisierung, habitualisierte Floskeln).

:: Codierte Sprache und Umwegkommunikation

Antisemitismus wird häufig verschleiert. Für die Erkennung – auch durch KI – ist es entscheidend, die typischen Muster codierter Sprache und von „Umwegkommunikation“ zu kennen:

- **Substitutionen:** „Zionisten“, „Globalisten“, „Israellobby“, „Bilderberger“.
- **Personalisierung:** Rothschilds, Soros als Chiffren für vermeintliche Macht.
- **Parolen und Idiome:** „From the river to the sea …“, „Auge um Auge“.
- **Kofferwörter und Neologismen:** „Satanyahu“, „Israhell“, „JewSA“; dazu Akronyme (NWO, ZOG) und Zahlencodes.

:: Israelbezogene Sprachmuster

Israelbezogener Antisemitismus äußert sich besonders über NS-Vergleiche, Dämonisierungen (Teufel, Satan), Krankheits-, Körper- und Tiermetaphern (Krebs, Pest, Fremdkörper) sowie Dysphemismen (Apartheid, Genozid, Regime). Für die quantitative wie qualitative Auswertung hilfreich ist zudem die Analyse semantischer Rollen: Wer wird als handelnd (Agens), wer als betroffen (Patiens) dargestellt? Welche Handlungsverben, welche Aktiv-/Passiv-Konstruktionen werden gewählt? Solche Perspektivierungen prägen die Wahrnehmung – etwa in der Medienberichterstattung.

Warum das für KI wichtig ist

Eine KI erkennt nur, was zuvor klar definiert wurde. Wer ein Sprachmodell zur Erkennung von Antisemitismus einsetzt, sollte im Falle der Nutzung eines Large Language Models (LLM) dem Modell eine fundierte Definition und konkrete Beispiele codierter Muster mitgeben oder das Modell nachtrainieren. Inhaltliche Expertise ist die Voraussetzung für eine sinnvolle technische Analyse – nicht ihr Ersatz.

6 · KI, Agenten und Large Language Models: Grundlagen

Large Language Models (LLMs) sind KI-Systeme, die darauf trainiert wurden, auf der Grundlage großer Textmengen Muster in Sprache zu erkennen und fortzuschreiben. Wichtig zu verstehen: Ein Sprachmodell rechnet nicht und „versteht“ nicht im menschlichen Sinne – es erkennt Muster und erzeugt daraus wahrscheinliche Fortsetzungen. Daraus ergeben sich sowohl die Stärken (schnelle Verarbeitung großer Sprachmengen) als auch die Grenzen (Fehler, Halluzinationen).

:: Datenarten und Datentypen

Für die Analyse ist die Unterscheidung der Datengrundlage zentral:

- **Strukturierte Daten:** z. B. CSV- oder Excel-Tabellen – maschinell gut auswertbar
- **Unstrukturierte Daten:** z. B. Fließtexte, Social-Media-Posts, PDF- und Word-Dokumente, Bilder, Videos.

:: Lokal oder in der Cloud?

LLMs lassen sich in der Cloud nutzen oder lokal auf eigenen Geräten betreiben. Lokale Modelle (etwa über Werkzeuge wie Ollama oder LM Studio) halten Daten auf dem eigenen Rechner und sind daher aus Datenschutzsicht oft vorzuziehen – sie sind allerdings rechen- und arbeitsspeicherintensiv. Künftig dürften immer mehr Modelle lokal laufen, wodurch die Datenverarbeitung stärker auf eigenen Geräten ermöglicht wird.

:: Vom reinen Sprachmodell zur „agentischen“ KI

Während frühe KI-Anwendungen reine Sprachmodelle waren, die zunächst nur auf eigene Quellen zugriffen, können moderne Systeme externe Quellen einbeziehen und „agentisch“ arbeiten: Sie führen eigene Funktionen, Aufgaben und Arbeitsschritte nacheinander aus und können ihre „eigenen Fähigkeiten“ sogar selbst programmieren. Das macht die Ergebnisse (zum Teil) nachvollziehbarer und überprüfbarer – ein wichtiges Kriterium für eine seriöse oder wissenschaftliche Analyse. Die agentischen Fähigkeiten ermöglichen eine tagelange Analyse und Durchführung verschiedener Schritte bis hin zu einer möglichen Lösung eines Problems.

:: Multimodale Modelle

Aktuelle Modelle sind zunehmend multimodal: Sie verarbeiten nicht nur Text, sondern auch Bilder, Audio und Video gemeinsam. Für die Antisemitismusanalyse ist das besonders relevant, weil sich Antisemitismus nur zu einem gewissen Teil auf der textuellen Ebene abspielt und in Zeiten von verstärktem Medienkonsum sozialer Medien und Emotionalisierung das Zusammenspiel von Bild, Ton, Musik und Videos eine noch stärkere Rolle einnimmt.

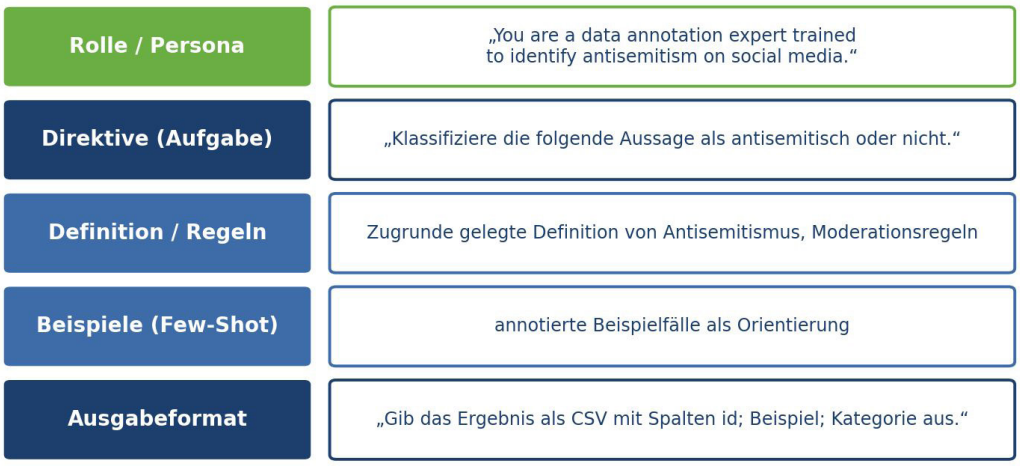
Forschung aus den letzten zwei Jahren zeigt: Spezialisierte, auf Aufgaben trainierte Vision-Language-Modelle erkennen solche Inhalte zunehmend gut, stoßen aber an Grenzen. Selbst Memes erschließen

sich erst aus subtilen Bild-Text-Bezügen und kulturellen Anspielungen; Modelle neigen zu Halluzinationen, zu vorschnellem „Über-Labeln“ als Hassrede und zu fehlender kultureller und sprachlicher Sensibilität – gerade außerhalb des englischsprachigen Raums. Menschliche Kontrolle bleibt daher häufig unverzichtbar.

7 · Wirksames Prompting für die Analyse

Ein Prompt ist der Input für ein generatives KI-Modell – meist Text, zunehmend aber auch Bild, Audio oder Video. Prompting ist ein iterativer Prozess: Man entwickelt und verfeinert die Eingabe Schritt für Schritt. Wer die Bausteine eines Prompts kennt, erzielt deutlich verlässlichere Ergebnisse. So können etwa Textpassagen automatisch klassifiziert werden. Eine solche Klassifizierung oder ein solches Antwortverhalten kann durch Fine-Tuning (nachträgliches Training entlang von klassifizierten Beispielen) verbessert werden. Alternativ können Beispiele bei der Eingabe (Prompt) – zum Beispiel im Benutzer:innen-Interface – mitgeliefert werden. Durch Automatisierungen in Form von kleinen Programmen oder durch Stapelverarbeitung (Batches) können so große Datenmengen automatisch analysiert werden.

Bausteine eines wirksamen Prompts



:: Zentrale Techniken

- **Zero-Shot:** die KI löst die Aufgabe ohne Beispiele – gut kombinierbar mit Rollen- und Stilvorgaben.
- **Few-Shot / In-Context-Learning:** wenige, ausgewogene Beispiele leiten das Modell an; oft genügen wenige, maximal etwa 20.
- **Chain-of-Thought:** das Modell artikuliert seine Überlegungen Schritt für Schritt („Let’s think step by step“).
- **Multilinguales Prompting:** englische Prompt-Templates sind teilweise wirksamer als rein deutschsprachige, da Modelle überwiegend mit englischen Daten trainiert werden.

:: Typische Fallstricke

- **Prompt-Sensitivität:** schon kleinste Änderungen (Leerzeichen, Reihenfolge) können die Ergebnisse stark verändern.
- **Overconfidence:** Modelle sind oft zu selbstsicher und neigen dazu, der Meinung der nutzenden Person zuzustimmen. Persönliche Wertungen gehören daher nicht in den Prompt.
- **Stereotype:** Modelle können Vorurteile aus ihren Trainingsdaten reproduzieren; gezielte Anweisungen zur Vorurteilsfreiheit („moral self-correction“) können gegensteuern.

8 · KI-gestützte Analyse: multimodal und kombiniert

In der Praxis verbinden sich Recherche, Strukturierung und Visualisierung zu einer Analyse-Pipeline. Bei einer Pipeline handelt es sich um nacheinander folgende Prozesse oder Arbeitsschritte. Man kann dabei vom Endprodukt her denken (etwa einem Netzwerkgraphen) oder einen Prozess als Abfolge automatisierter KI-Schritte gestalten – zunächst eine Tabelle erzeugen lassen, darauf aufbauend eine Visualisierung.

Eine KI-gestützte Analyse-Pipeline

:: Strukturierte Ergebnisse erzeugen

Ein praktischer Kniff aus den Workshops: Man lässt die KI Analyseergebnisse direkt als strukturierte Tabelle ausgeben, zum Beispiel mit der Aufforderung, eine CSV-Datei (mit Semikolon getrennt) mit den Spalten id; Beispiel; Überkategorie; Realisierung; antisemitisches Muster zu erstellen. So lassen sich Beispiele systematisch kategorisieren und anschließend in einer Tabellenkalkulation weiterverarbeiten.

:: Quantitativ und qualitativ kombinieren

Für die sprachwissenschaftliche Auswertung eignen sich Korpus-Werkzeuge wie AntConc oder andere Programme (z. B. für Cluster, N-Gramme, Wordclouds) – ergänzend oder alternativ zur KI. Multimodale Auswertungen durch andere Programme und Cloudanwendungen ermöglichen es, auch Bilder und Videostreams einzubeziehen, etwa durch die Analyse einzelner Frames in festgelegten Intervallen in Kombination mit Texterkennung (OCR) oder durch Modelle, die ganze Videos analysieren.

:: Beispiel: KI-gestützte Netzwerkanalyse und Retrieval-Augmented Generation (RAG)

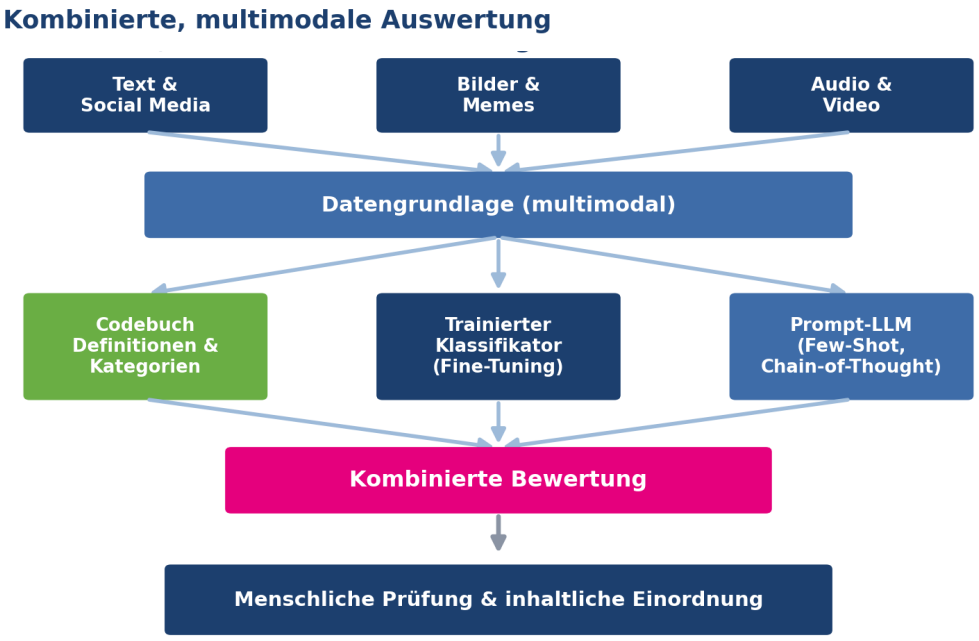
Netzwerke von Sinnzusammenhängen oder Akteur:innen lassen sich in der Forschung sichtbar machen, indem man zunächst eine strukturierte Tabelle erstellt – etwa nach Sinnzusammenhängen oder Begriffen mit bestimmter Zugehörigkeit und Quellen – und daraus mit Werkzeugen einen Netzwerkgraphen mit Knoten (nodes) und Kanten (edges) erzeugt. Für große Datenmengen ist es ressourcenschonend, vorab per Schlagwörtern oder über Retrieval-Augmented Generation (RAG) zu filtern.

:: Kombinierte Auswertung: Codebücher, trainierte Klassifikatoren und Prompt-LLMs

Kein einzelner Ansatz ist für sich genommen optimal. In der aktuellen Forschung wird daher eine kombinierte Auswertung und Analyse diskutiert, die unterschiedliche Verfahren auf eine gemeinsame, möglichst multimodale Datengrundlage anwendet und ihre Ergebnisse zusammenführt:

- **Codebücher als Kontext:** Ein Codebuch hält Definitionen, Kategorien und Beispiele fest – ein mögliches Fundament einer traditionellen Inhaltsanalyse. Gibt man dieses Codebuch der KI als Kontext mit, kann sie automatische Klassifikationen entlang nachvollziehbarer, fachlich begründeter Kriterien vornehmen. Studien zu „Codebook-LLMs“ zeigen, dass Definitionen und Kontext die Qualität deutlich verbessern.
- **Trainierte Klassifikatoren:** Auf konkrete Aufgaben feinabgestimmte (fine-getunte) Modelle – etwa auf Basis von BERT-Varianten oder mit LoRA-Adaptern – liefern bei ausreichend Trainingsdaten oft relativ zuverlässige Ergebnisse für wiederkehrende Klassifikationen.
- **Prompt-LLMs:** Aktuelle, per Prompt gesteuerte Sprachmodelle (Zero-/Few-Shot, Chain-of-Thought) sind flexibel, schnell einsetzbar und gut für neue oder seltene Phänomene – allerdings sensibler gegenüber Formulierungen.

Die Stärke liegt in der Kombination: Ein Codebuch sorgt für fachliche Konsistenz, ein trainierter Klassifikator für Zuverlässigkeit bei großen Mengen, ein Prompt-LLM für Flexibilität. Eine zusammengeführte Bewertung kann abschließend menschlich geprüft und eingeordnet werden.



Hinweis aus der Forschung

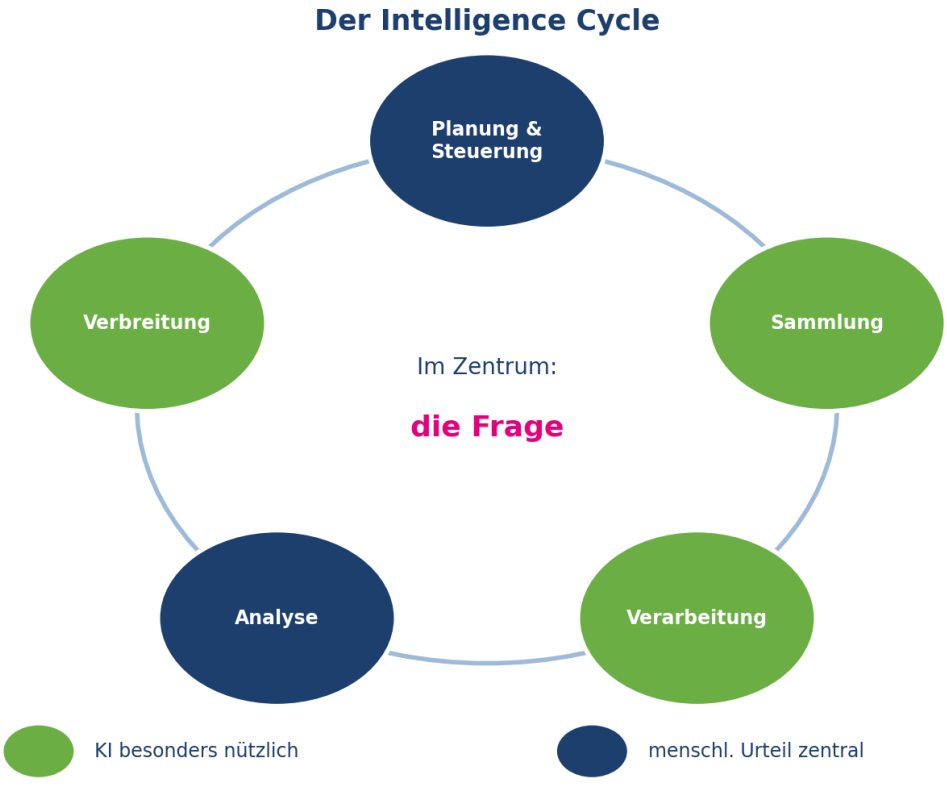
„LLM-Hacking"-Risiko: Untersuchungen zur LLM-gestützten Annotation warnen, dass Ergebnisse stark vom Modell und von Details des Prompts abhängen. Wer KI zur Klassifikation einsetzt, sollte die Verfahren dokumentieren, an von Menschen annotierten Beispielen validieren und Ergebnisse stichprobenartig prüfen.

9 · Journalismus, Recherche und Open Source Intelligence

OSINT (Open Source Intelligence) bezeichnet unter anderem im Journalismus Erkenntnisse, die aus öffentlich verfügbaren Informationen gewonnen werden. Im Zentrum jeder Recherche und Analyse steht eine klare Frage – ohne sie verrennt man sich schnell. Datenschutzrechtliche und andere rechtliche Aspekte müssen dabei vorab geklärt werden.

:: Der Intelligence Cycle

Strukturierte Recherche folgt einem Kreislauf. Künstliche Intelligenz ist dabei nicht überall gleich nützlich: Bei der Fragestellung und der Analyse braucht es menschliche Kreativität und kritisches Urteil, während KI bei Sammlung, Verarbeitung und Verbreitung wertvolle Dienste leistet. Mustererkennung und Visualisierung sind möglich – fundierte Schlussfolgerungen müssen jedoch kritisch geprüft werden.



:: Wo Informationen liegen – und wie man sie bewertet

Quellen müssen kritisch bewertet werden – nach Glaubwürdigkeit und Herkunft, Aktualität und Nachvollziehbarkeit, Kontext und möglichen Manipulationen sowie nach ethischen und rechtlichen Rahmenbedingungen. Gerade in Konflikten wie Israel-Gaza werden – teils unbewusst – falsche oder veraltete Informationen verbreitet. Hilfreich sind Werkzeuge zur Beweissicherung und -archivierung sowie zur Reverse-Bildsuche und Metadatenanalyse.

:: Sicher recherchieren: OPSEC

Operational Security ist immer kontextabhängig: Vor wem will ich mich schützen? Ein VPN allein schützt nur begrenzt – der VPN-Anbieter selbst kennt die Aktivität. Je nach Schutzbedarf sind ein TOR-Browser, eine virtuelle Maschine oder ein eigenes Recherchegerät sinnvoll.

10 · KI für Kommunikation, Kampagnen und Berichte

KI kann die Kommunikationsarbeit zivilgesellschaftlicher Organisationen spürbar unterstützen und vereinfachen. In der Öffentlichkeitsarbeit lässt sich KI praxisnah einsetzen:

- Texte für Website und Social Media schneller entwerfen.
- Veranstaltungen und interne Abläufe mithilfe von KI-Lösungen planen.
- Einen Social-Media-Redaktionsplan mit KI erstellen.
- Social-Media-Kampagnen gemeinsam mit Kooperationspartner:innen entwickeln.

:: Automatisierung von Analysen und Schreibprozessen

Auch die Berichterstellung lässt sich automatisieren – etwa, indem eine Rede oder ein Vortrag aufgenommen, automatisiert verschriftlicht, zusammengefasst und für einen Artikel aufbereitet wird. Agentische Systeme erledigen solche Aufgaben in nachvollziehbaren Schritten. Entscheidend bleiben die menschliche Endredaktion und die Richtigkeit der Inhalte: Wer Ergebnisse veröffentlicht, haftet für deren Richtigkeit. Vor der Veröffentlichung ist daher stets zu prüfen, ob die Informationen zutreffen. Zudem müssen rechtliche Kennzeichnungs- und andere Pflichten berücksichtigt werden. Auch gelten unterschiedliche Leitlinien und Empfehlungen für verschiedene Berufsgruppen.

Verbreitung adressatengerecht gestalten
Ob Bericht, Briefing oder Dashboard – die Aufbereitung der Ergebnisse muss adressatengerecht, klar und präzise erfolgen, damit Erkenntnisse zeitnah und verständlich für Entscheidungen genutzt werden können.

11 · Autonome KI-Agenten: Werkzeug und Risiko

Eine neuere Entwicklungsstufe sind autonome KI-Agenten: Systeme, die ein Ziel nicht nur in einem Schritt beantworten, sondern eigenständig planen, Werkzeuge nutzen (Browser, Programmcode, Schnittstellen), die Ergebnisse beobachten und ihr Vorgehen wiederholt anpassen – bis die Aufgabe erledigt ist. Manche Agenten entwickeln dabei eigene „Skills“ weiter und werden mit der Zeit besser.

In den Jahren 2025 und 2026 ist eine Vielzahl solcher Agenten entstanden – darunter OpenHands (autonome Softwareentwicklung), Manus (ein „autonomer digitaler Mitarbeiter“), die lokal lauffähigen Agenten Hermes und OpenClaw sowie Claude CoWork und OpenAIs Operator. Sie bearbeiten teilweise tagelang Aufgabenstellungen (auf Servern oder lokal), werden über Chat- und andere Apps gesteuert, interpretieren Screenshots, klicken, tippen und ketten mehrstufige Aufgaben zusammen – vom Erstellen eines Entwurfes eines wissenschaftlichen Buches über das Extrahieren und Speichern von Daten in Datenbanken bis zum Versand von zielgruppenspezifischen E-Mails an einen großen Personenkreis. Sehr komplexe aufeinanderfolgende Aufgaben können dabei abgearbeitet werden.

:: Chance für die Arbeit gegen Antisemitismus

Für zivilgesellschaftliche Organisationen können Agenten wiederkehrende Abläufe übernehmen: Quellen automatisiert sammeln und vorstrukturieren, Monitoring-Routinen ausführen, Berichte vorbereiten. Richtig eingesetzt – mit klarer Aufgabenstellung, nachvollziehbaren Schritten und menschlicher Endkontrolle – können sie knappe Personalressourcen entlasten.

:: Risiko der Automatisierung von Hass

Dieselbe Autonomie kann missbraucht werden. Agenten können Desinformationskampagnen weitgehend selbstständig betreiben: Inhalte erzeugen, Accounts bespielen, auf Gegenrede reagieren – rund um die Uhr und in großem Maßstab. Je leistungsfähiger und je weniger kontrolliert die dafür genutzten LLMs sind, desto größer ist das Missbrauchspotenzial. Autonome Agenten verschärfen damit die bereits beschriebene Bedrohung durch Einflussoperationen. Eine besondere Herausforderung: Mehrere Agenten und LLMs können dabei kombiniert werden und so Stärken und Schwächen (z. B. bei der Verhinderung von Hassnachrichten) gezielt genutzt werden.

12 · Chancen, Grenzen und Verantwortung generativer KI

Künstliche Intelligenz ist kein neutrales Instrument, sondern basiert auf Daten und Modellen, die von Menschen entwickelt werden. Generative KI, also die Nutzung von ChatGPT, der Chatfunktionen von Claude, Mistral AI oder anderen europäischen Modellen, stellt inzwischen nur noch einen Teil der Nutzung von KI dar. Dennoch bleiben einige Rahmenbedingungen zentral: Fehlerhafte Datensätze oder unausgewogene Trainingsgrundlagen können zu Fehlentscheidungen und der Verbreitung von Fehlinformationen führen. Transparenz, Nachvollziehbarkeit und die Einhaltung rechtlicher Rahmenbedingungen sind Voraussetzungen für einen verantwortungsvollen Einsatz.

:: Wo generative KI hilft – und wo nicht

Phase	Eignung von KI
Fragestellung	eher menschlich – braucht Kreativität und Kontext; KI-Unterstützung möglich
Sammlung	gut geeignet – Crawling, Scraping, Monitoring
Verarbeitung	gut geeignet – Filtern, Bereinigen, Strukturieren
Eigen-Analyse	teilweise – Mustererkennung und Analyse von großen Datenmengen ja, Schlussfolgerungen kritisch prüfen
Verbreitung	gut geeignet – adressatengerechte Aufbereitung

:: Leitlinien für den verantwortungsvollen Einsatz

- **Mensch im Zentrum:** KI unterstützt, ersetzt aber nicht die fachliche Bewertung. Endredaktion und Verantwortung bleiben menschlich.
- **Quellen prüfen:** Halluzinationen und Falschinformationen einkalkulieren; Ergebnisse gegenprüfen.
- **Prävention von Hass und Meinungsfreiheit:** Moderationssysteme müssen zwischen Hassrede und zulässiger Meinungsäußerung unterscheiden. Der Kampf gegen Antisemitismus muss die Verminderung der Verbreitung von Antisemitismus durch Social Media und KI-Plattformanbieter beinhalten.

13 · Ausblick

Die technologische Entwicklung schreitet rasch voran. Künftig werden immer mehr Tätigkeiten durch KI-Agenten in der Cloud oder lokal durchgeführt. Die Stärkung lokaler Modelle und die Datenverarbeitung auf eigenen Geräten stärken den Datenschutz. Agentische Systeme werden Arbeitsschritte zunehmend eigenständig und nur zum Teil nachvollziehbar ausführen. Zugleich zeichnen sich neue Herausforderungen ab: KI-gestützte Desinformation und Fake News werden zunehmen, und der Fortschritt im Quantum-Computing wird absehbar bisherige Verschlüsselungstechniken in Frage stellen. Agenten werden zunehmend für Angriffe auf die IT-Infrastruktur, staatliche Behörden, politische Akteur:innen, Hochschulen und Zivilgesellschaft genutzt werden.

Für die Zivilgesellschaft, Wissenschaft und Politik bedeutet das: Wissen muss kontinuierlich aktualisiert, Methoden müssen erprobt und Erfahrungen geteilt werden. Antisemitismus findet immer mehr auch im Netz statt; wer ihm wirksam begegnen will, muss die digitalen Werkzeuge beherrschen – und verantwortungsvoll einsetzen.

KI ist ein Werkzeug. Ob es der Demokratie nützt oder schadet, entscheidet sich daran, wer es wie einsetzt – mit Fachwissen, kritischem Urteil und Verantwortung.

Über das IIBSA und das DigiLabs

Das Internationale Institut für Bildung, Sozial- und Antisemitismusforschung e.V. (IIBSA) wurde 2006 gegründet und ist seither in der Erforschung und Bekämpfung von Antisemitismus aktiv – mit einem Schwerpunkt auf Digitalisierung, Automatisierung und Künstlicher Intelligenz. Das Institut organisiert Konferenzen, Workshops und Vorträge, veröffentlicht wissenschaftliche Arbeiten und berät Ministerien, Politik, Journalist:innen, Kommunen, jüdische Gemeinden und zivilgesellschaftliche Institutionen.

Diese Handreichung entstand im Rahmen des Projekts „DigiLabs – Digitalisierung und Künstliche Intelligenz gegen Antisemitismus“. In zehn Workshops und der Etablierung eines Austausch-Forums brachte das Projekt zivilgesellschaftliche Akteur:innen, jüdische Organisationen und Expert:innen aus Wissenschaft, Recht und Cybersicherheit zusammen, um die Nutzung von KI in der Analyse und Bekämpfung von Antisemitismus zu erproben und gemeinsam weiterzuentwickeln.

:: Kontakt

Internationales Institut für Bildung, Sozial- und Antisemitismusforschung e.V. (IIBSA)
Pariser Platz 6a, 10117 Berlin · Telefon: 030 55 214 934
mail@iibsa.org · www.iibsa.org

:: Presseanfragen

Presseanfragen richten Sie bitte an: presse@iibsa.org

Gefördert durch:

Beauftragter der Bundesregierung
für jüdisches Leben in Deutschland
und den Kampf gegen Antisemitismus

aufgrund eines Beschlusses
des Deutschen Bundestages