DIW Wochenbericht 29+30 2022

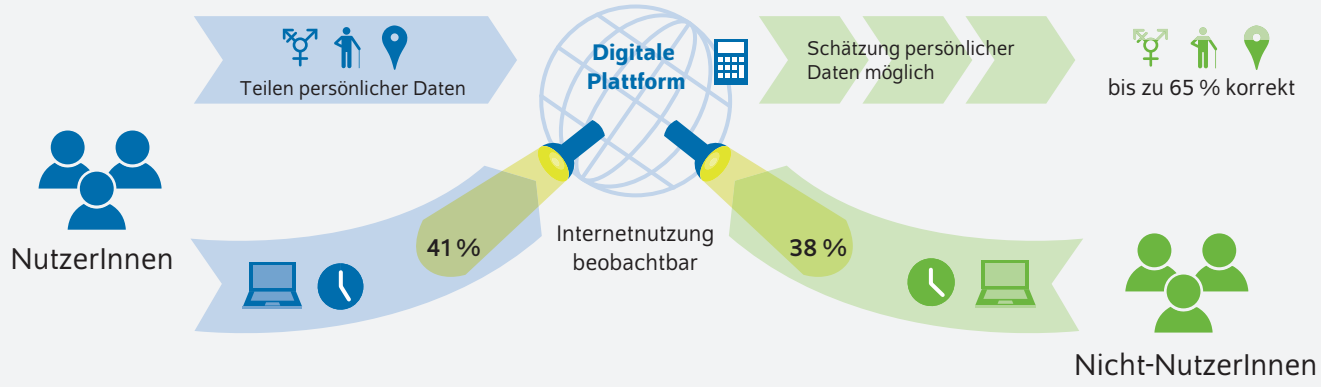
AUF EINEN BLICK

Plattformen wie Facebook können mehr als die Hälfte der Internetaktivität beobachten

Von Hannes Ullrich, Christian Peukert, Maximilian Schäfer und Luis Aguiar

- Online-Plattformen wie Facebook sind technisch in der Lage, mithilfe von Trackern das Surfverhalten von InternetnutzerInnen zu verfolgen
- Anbieter können Like-, Share- oder Login-Buttons nutzen, um von NutzerInnen besuchte Seiten nachzuvollziehen und zu speichern
- Schätzungen zeigen: Bis zu 52 Prozent der Aktivität im Internet kann theoretisch ausgelesen werden
- Plattformen können Surfverhalten ihrer NutzerInnen auswerten und dadurch Rückschlüsse auf Personen ziehen, die die Plattform selbst nicht nutzen
- Transparente Nutzeridentifikatoren und unabhängige Datenbroker könnten mehr Datenschutz etablieren

Große digitale Plattformen wie Facebook können Surfverhalten beobachten und daraus Schattenprofile erstellen



Quelle: Eigene Darstellung.

ZITAT

„Unsere Berechnungen zeigen, dass Plattformen wie Facebook die technischen Möglichkeiten haben, die Internetaktivitäten und Eigenschaften vieler Menschen in erheblichem Ausmaß nachzuvollziehen. Damit Regulierungen unter Abwägung von Nutzen und Risiken der Datennutzung umgesetzt werden können, sollten die europäischen Aufsichtsbehörden gestärkt werden.“ — ***Hannes Ullrich*** —

MEDIATHEK

Audio-Interview mit Hannes Ullrich
www.diw.de/mediathek

Plattformen wie Facebook können mehr als die Hälfte der Internetaktivität beobachten

Von Hannes Ullrich, Christian Peukert, Maximilian Schäfer und Luis Aguiar

ABSTRACT

Große digitale Plattformen wie Amazon, Apple, Facebook, Google und Microsoft haben die technischen Möglichkeiten, umfangreiche Daten ihrer NutzerInnen zu sammeln. Beim Surfen im Internet kann dies durch das Laden kleinerer Programme und Identifikatoren wie Cookies geschehen, die das Beobachten von Besuchen auf Webseiten ermöglichen. Gemäß vorliegender Schätzung könnte Facebook im Durchschnitt 40 Prozent der im Internet verbrachten Zeit beobachten – weitestgehend unabhängig von den demografischen Eigenschaften der NutzerInnen und den Arten besuchter Webseiten. BesucherInnen von Webseiten teils sensibler Kategorien wie Wettspielseiten, Seiten mit Gesundheits- und Ernährungsinformationen oder Immobiliengeschäften sind mit 60 bis 80 Prozent möglicher beobachteter Zeit besonders betroffen, während Email- oder Messenger-Dienste mit unter 20 Prozent weniger betroffen sind. Da Facebook-NutzerInnen auch persönliche Eigenschaften mit Facebook teilen, die mit dem Surfverhalten außerhalb von Facebook korrelieren – beispielsweise Alter und Geschlecht –, kann Facebook die persönlichen Eigenschaften selbst von NutzerInnen, die sich nie bei Facebook angemeldet haben, zum Erstellen von Schattenprofilen herleiten. In der Studie können so persönliche Eigenschaften von Facebook-NutzerInnen zu 60 bis 80 Prozent korrekt geschätzt werden. Für Nicht-NutzerInnen von Facebook gelingt dies zwar nur zu rund 60 Prozent korrekt, dennoch zeigt dies, dass das Surfverhalten Informationen auch über diese Personengruppe enthält. So könnte Facebook Schattenprofile nutzen, um auch außerhalb der eigenen Plattform zielgerichtete Werbung zu vermarkten.

Für Online-Plattformen sind Anzahl und Vielfalt ihrer NutzerInnen essentiell für den Erfolg: Treffen sich Anbieter und potentielle AbnehmerInnen von Produkten auf Plattformen, so profitieren beide Seiten davon, wenn ihnen ein größeres Angebot an Anbietern und AbnehmerInnen gegenübersteht. Auf diesem Geschäftsmodell basiert der Erfolg von großen Plattformen wie Amazon im Einzelhandel, Facebook in den sozialen Netzwerken, Google in der Internetsuche und Apple sowie Google mit App Stores für mobile Endgeräte.

Im Geschäftsmodell von Plattformen spielen Werbeeinahmen eine zentrale Rolle.[1] Für Google und Meta, dem Mutterkonzern von Facebook, zusammen beliefen sich diese im Jahr 2021 auf 324 Milliarden US-Dollar.[2] Dass die Werbeeinnahmen so hoch ausfallen, liegt daran, dass digitale Plattformen durch ihre Größe einen einzigartigen Zugang zu einer Vielzahl von KonsumentInnen haben und damit auch zu deren Aufmerksamkeit und Zeit.

Weil die NutzerInnen teils weitreichende persönliche Informationen mit Plattformen wie Facebook teilen, können daraus abgeleitete individuelle Interessen als Grundlage für personalisierte Werbung genutzt werden. Eine einfache, gängige Praxis – auch außerhalb von Facebook – ist das Erstellen demografischer Profile, bei denen Personen in Gruppen nach Alter, Geschlecht oder Wohnort unterteilt werden. Anzeigen werden dann für gewünschte demografische Profile verkauft.[3] Hiervon profitieren potentiell auch KonsumentInnen, da ihnen Produkte angeboten werden, die ihren Präferenzen besser entsprechen. Gleichzeitig gibt es aber auch Bedenken, dass die Privatsphäre der NutzerInnen verletzt wird und Möglichkeiten individueller Preisgestaltung zu Ungunsten von KonsumentInnen entstehen, wenn Plattformen detaillierte Informationen über einzelne NutzerInnen erstellen können.

1 Vgl. Competition and Markets Authority (2020): Online platforms and digital advertising market study (online verfügbar, abgerufen am 24. Juni 2022. Dies gilt für alle anderen Online-Quellen dieses Berichts, sofern nicht anders vermerkt).

2 Vgl. die Darstellung von Statista (online verfügbar).

3 Vgl. Nico Neumann, Catherine E. Tucker und Timothy Whitfield (2019): Frontiers: How effective is third-party consumer profiling? Evidence from field studies. Marketing Science 38 (6), 918–926.

DOI: https://doi.org/10.18723/diw_wb:2022-29-1

Die Studie, auf der dieser Bericht basiert, zeigt, in welchem Ausmaß eine Plattform wie Facebook die technische Möglichkeit hat, die Inhalte, die im Internet konsumiert werden, zu beobachten und zu nutzen.[4] Grundlage für die Analyse bilden Daten zu allen Internetseiten – insgesamt 18,17 Millionen Klicks sowie die auf jeder Webseite verbrachte Zeit –, die im Jahr 2016 von einer repräsentativen Stichprobe von 4 853 InternetnutzerInnen in den USA besucht wurden (Kasten 1). In der Studie werden die Internet-Nutzungsdaten mit archivierten Informationen möglicher Datenübertragungen zu Facebook auf Ebene einzelner Domains verknüpft, um zu untersuchen, welcher Anteil der Surf-Aktivitäten verschiedener NutzerInnen im Internet technisch von Facebook gesehen werden kann. Schließlich zeigt die Studie quantitativ, wie persönliche Informationen in Form von Schattenprofilen auch ohne explizites Wissen vieler InternetnutzerInnen anhand von Internet-Nutzungsdaten und Methoden des maschinellen Lernens erlangt werden können.

Plattformen können Surfverhalten auch auf externen Seiten beobachten

Beim täglichen Surfen im Internet wird ein individueller Fußabdruck hinterlassen. Dieser Fußabdruck ist eine Abfolge unzähliger besuchter Webseiten, vom Wetterdienst, über Sportseiten, Geschäfte für Kleidung, Technik oder Spielzeug, hin zu Internetforen, Unterhaltungsseiten und Sachinformationen zum Gärtnern oder Kochen oder zu medizinischen Informationen.

Einen Teil dieses Fußabdrucks können Plattformen wie Twitter, LinkedIn, Facebook sowie das zum Facebook-Konzern Meta gehörende Instagram durch Engagement-Buttons beobachten, speichern und weiterverwenden. Im Fall von Facebook können dies Like- und Share-Buttons sein, aber auch Login-Hilfen können so genutzt werden.[5] Hierbei spielt es keine Rolle, ob BesucherInnen einer Webseite Facebook-NutzerInnen sind oder nicht – und auch nicht, ob sie auf die entsprechenden Buttons klicken oder nicht. Sobald eine Website aufgerufen wird, können externe Anbieter durch das Laden eines Buttons die Nutzeraktivität auf der Webseite erkennen.[6] Die gängigste Form der Datensammlung mit dem Ziel, Konsumentenprofile zu erstellen, ist das Tracking anhand von Third-Party-Cookies – kleine Textdateien, die individuelle NutzerInnen identifizieren (Kasten 2). Für Unternehmen, denen personalisierte Werbung höhere Einnahmen in Aussicht stellt, bestehen hohe Anreize,

Kasten 1

Daten und Analyse

Die Studie nutzt Daten einer durch die Marktforschungsfirma Nielsen erstellten repräsentativen Stichprobe von 4 853 InternetnutzerInnen für das Jahr 2016. In der Erhebung haben die Befragten persönliche Merkmale angegeben und ihr Surfverhalten auf einem Desktop-Computer für ein Jahr transparent gemacht. Das Surfverhalten wird anhand des gesamten Verlaufs aller von den individuellen NutzerInnen besuchten Webseiten gemessen. Hierbei werden sowohl die Anzahl der Clicks je Website als auch die Verweildauer auf der entsprechenden Seite aufgenommen. Für jede besuchte Webseite wird aus einem zweiten Datensatz der Initiative *httparchive* hinzugespielt, welche externen Dienstleister beim Laden der Webseite Daten übertragen konnten. Durch das Verknüpfen dieser beiden Datenquellen kann der durch einen externen Dienstleister beobachtbare Anteil des Surfverhaltens für alle InternetnutzerInnen quantifiziert werden. Besucht eine Person die Facebook-Plattform im Verlauf eines Jahres nicht, wird angenommen, dass diese Person Facebook nicht nutzt. So wird zwischen NutzerInnen und Nicht-NutzerInnen von Facebook unterschieden.

Durch Klassifizierungsmethoden kann die Verbindung der in den Daten enthaltenen persönlichen Eigenschaften zu dem Surfverhalten hergestellt werden. Besteht ein systematischer Zusammenhang, können Datenexternalitäten entstehen. Das bedeutet, dass Nicht-NutzerInnen von Facebook Kosten oder Nutzen dadurch tragen, dass Facebook-NutzerInnen durch das Teilen ihrer Daten der Plattform erlauben, diese Zusammenhänge zwischen Surfverhalten und persönlichen Eigenschaften zu messen.

In der Studie wird eine gängige Methode des maschinellen Lernens verwendet, bei der eine Vielzahl einzelner Klassifizierungsmethoden kombiniert werden, um deren individuelle Schwächen zu reduzieren. Das hier verwendete *Extreme Gradient Boosting*[1] bildet eine Ansammlung von Klassifikationsbäumen, um für ein vorgegebenes Kriterium die beste Balance zwischen präziser und unverzerrter Klassifikation zu finden. Insbesondere wird das Klassifizierungsmodell nur anhand von Daten eines Teils der Facebook-NutzerInnen erstellt. Dann quantifiziert die Studie die Schätzgenauigkeit der Eigenschaften der ausgelassenen rund 1000 NutzerInnen sowie der rund 1000 Nicht-NutzerInnen von Facebook. Die Qualität der Schätzungen persönlicher Eigenschaften wird durch die Korrektklassifikationsrate gemessen. Diese gibt an, welchen Anteil an allen Beobachtungen ein Algorithmus korrekt klassifiziert.

1 Vgl. Trevor Hastie, Robert Tibshirani und Jerome H. Friedman (2009): The elements of statistical learning: data mining, inference, and prediction. New York; Tianqi Chen und Carlos Guestrin (2016): Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794.

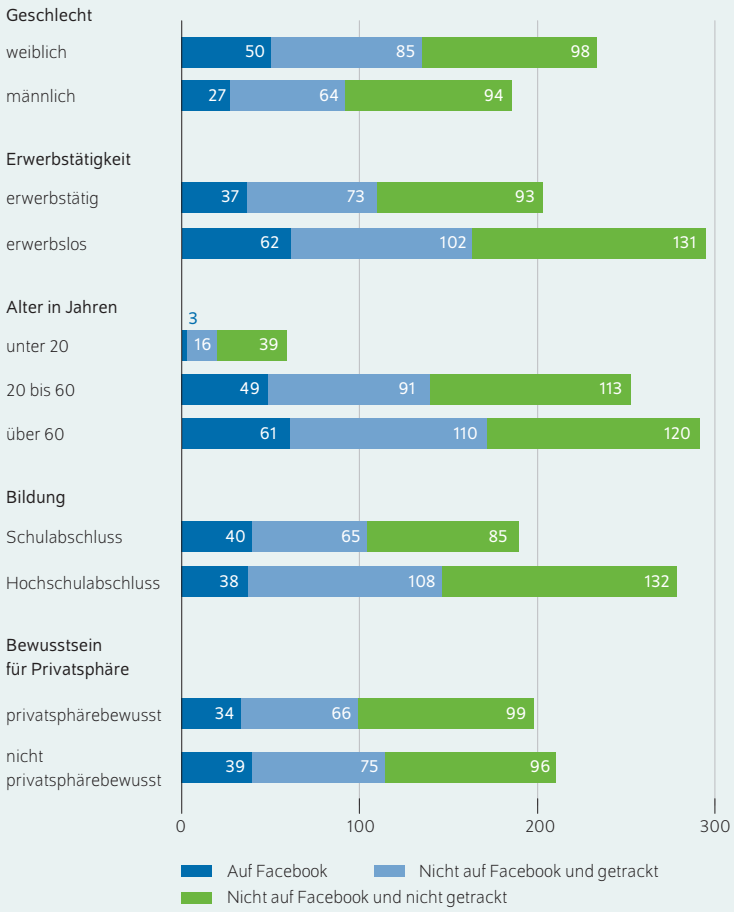
4 Vgl. Luis Aguiar, Christian Peukert, Maximilian Schäfer und Hannes Ullrich (2022): Facebook Shadow Profiles. arXiv Working Paper (online verfügbar).

5 Hierbei ist anzumerken, dass weitreichendes Tracking auch durch andere Dienste wie Google, das zu Alphabet gehört, betrieben werden kann, ohne auf sichtbare Engagement-Buttons zurückzugreifen, vgl. Sebastian Schelter und Jérôme Kunegis (2018): On the ubiquity of web tracking: Insights from a billion-page web crawl. The Journal of Web Science 4. Zur technischen Schwierigkeit, in der bestehenden Struktur der Online-Werbungsindustrie Datensicherheit zu gewährleisten, vgl. die Webseite des Irish Council for Civil Liberties.

6 Manche Browser und deren Erweiterungen erlauben das automatische Löschen von Cookies oder unterdrücken das Laden von Werbung und Trackern. Der prominenteste Browser ist hier Firefox, den allerdings nur rund drei Prozent der InternetnutzerInnen verwenden. Vgl. die Übersicht bei Statcounter (online verfügbar).

Abbildung 1

Beobachtbare im Internet verbrachte Zeit
In Stunden



Anmerkungen: Daten aus 2016.

Quelle: Nielsen Clickstream-Daten, httparchive Tracking-Daten, eigene Berechnungen.

Plattformen wie Facebook können über alle demografischen Gruppen hinweg in ähnlichem Maß die im Internet verbrachte Zeit beobachten.

Kasten 2

Datenerhebung im Internet

Das World Wide Web Consortium definiert Tracking als „das Sammeln von Daten über die Aktivitäten eines bestimmten Nutzers in mehreren unterschiedlichen Kontexten sowie die Speicherung, Nutzung oder Weitergabe von Daten, die aus diesen Aktivitäten abgeleitet wurden, außerhalb des Kontexts, in dem sie stattfanden."[1] Gängige Kontexte im Internet sind Webseiten und Apps.

Um Daten über Kontexte hinweg zu verknüpfen, werden Cookies als Identifikatoren verwendet. Dies sind kleine Textdateien, die auf dem Endgerät der NutzerInnen gespeichert werden. Bei First-Party-Cookies geschieht dies durch den Anbieter der besuchten Webseite. Werden diese von Drittanbietern geladen und Daten außerhalb der besuchten Webseite verknüpft, so spricht man von Third-Party-Cookies. Datenbroker, die Daten sammeln und zusammenführen, haben sich dank Tracking-Technologien zu einem bedeutenden Wirtschaftszweig entwickelt. Sie verkaufen Konsumentenprofile zum Einsatz zielgerichteter Online-Werbung.[2]

Die am 25. Mai 2018 implementierte Datenschutz-Grundverordnung (DSGVO) regelt die Verarbeitung personenbezogener Daten nach sechs Grundsätzen. Die bekannten Abfragen der Zustimmung zur Datenverarbeitung leiten sich aus dem Grundsatz der Rechtmäßigkeit ab. Sie bieten Plattformen einen einfachen Weg, die Rechtmäßigkeit weitgehender Datenverarbeitung nachzuweisen. Weitere wichtige Grundsätze sind die Datenminimierung und die Zweckbindung der Datenverarbeitung. Der sich in der Rechtssetzung befindende europäische Digital Markets Act (DMA) soll diese Prinzipien für dominante Plattformen – sogenannte Gatekeeper – noch weiter stärken.

1 Vgl. die Webseite des World Wide Web Consortium.

2 Vgl. Federal Trade Commission (2014): Data Brokers: A Call for Transparency and Accountability. Washington, DC (online verfügbar).

kohärente Nutzerprofile über Kontexte wie Anwendungen, Webseiten oder Endgeräte hinweg zu erstellen.[7]

Zwar entschied der Europäische Gerichtshof im Jahr 2019, dass BetreiberInnen von Webseiten gemäß Datenschutz-Grundverordnung (DSGVO) die Zustimmung für die Datenübertragung anhand von Engagement-Buttons einholen müssen, allerdings zeigt sich, dass NutzerInnen wenig Kenntnis der Tracking-Praxis im Internet haben und nur wenige aktiv die Zustimmung zur Datennutzung verweigern.[8] Auch rechtlich stellt die Art und Weise, wie AnbieterInnen von Webseiten Zustimmung einholen und speichern, einen noch offenen Streitpunkt in der Umsetzung der DSGVO dar.[9] In den vergangenen Jahren hat sich das Ausmaß der Datensammlung und -nutzung insbesondere durch Facebook in verschiedenen Fällen offenbart.[10] Zeitgleich wurde auch die Rolle von Datenbrokern deutlich, die umfangreiche Nutzerdaten zusammenführen und mit diesen handeln.[11] Gleichwohl führt die Verschwiegenheit der Plattformen bezüglich ihrer Datennutzung

7 Vgl. Tesary Lin und Sanjog Misra (2022): Frontiers: The Identity Fragmentation Bias. Marketing Science, 41 (3).

8 Vgl. Arunesh Mathur et al. (2018): Characterizing the use of browser-based blocking extensions to prevent online tracking. Fourteenth Symposium on Usable Privacy and Security, 103–116; Garrett A. Johnson, Scott K. Shriver und Shaoyin Du (2020): Consumer privacy choice in online advertising: Who opts out and at what cost to industry? Marketing Science 39 (1), 33–51.

9 Vgl. Entscheidung der belgischen Datenschutzbehörde vom 2. Februar 2022 (online verfügbar).

10 Vgl. die Entscheidung der Federal Trade Commission vom 18. Dezember 2019 (online verfügbar); Samantha Murphy Kelly und Clare Duffy (2021): Facebook whistleblower testifies company 'is operating in the shadows, hiding its research from public scrutiny'. CNN vom 6. Oktober (online verfügbar).

11 Vgl. Steven Melendez (2019): A landmark Vermont law nudges over 120 data brokers out of the shadows. Fastcompany vom 2. März (online verfügbar).

und der in der Datenauswertung verwendeten Algorithmen zu mangelnder quantitativer Evidenz über das Ausmaß des Trackings sowie dem dadurch gewonnenen Nutzen für Plattformen und deren NutzerInnen.

Für die in der Studie analysierte repräsentative Stichprobe hätte Facebook die Möglichkeit, NutzerInnen auf im Durchschnitt 52 Prozent der von ihnen besuchten Internetseiten zu sehen. Diese Webseiten entsprechen 40 Prozent der Zeit, die die analysierten NutzerInnen im Internet verbrachten.

Zwar unterscheiden sich Personengruppen in der Studie in der Intensität der Internetnutzung und in der Zeit, die sie auf der Facebook-Plattform verbringen (Abbildung 1), aber der Anteil des beobachteten Surfverhaltens ist über alle demografischen Personengruppen stabil. Die durch Facebook beobachtbare Internetnutzung außerhalb der Facebook-Plattform ist für alle Nutzergruppen ungefähr doppelt so hoch wie auf der Plattform selbst.

Ein naheliegender Gedanke ist, dass NutzerInnen die Facebook-Plattform meiden, um das Ausmaß, zu dem Facebook das eigene Surfverhalten beobachten kann, zu verringern. Die Studie zeigt jedoch, dass dies nicht einfach möglich ist. Ein Vergleich von aktiven Facebook-NutzerInnen und NutzerInnen, die überhaupt keine Zeit auf der Facebook-Plattform verbringen, zeigt kaum Unterschiede (Abbildung 2). Plattformen wie Facebook könnten 41 Prozent der online verbrachten Zeit von NutzerInnen der Plattform beobachten – und 38 Prozent bei Nicht-NutzerInnen. Damit ist die Variation zwischen Personen, die Facebook nutzen, und Personen, die Facebook nicht nutzen, geringer als die Variation über demografische Gruppen hinweg.

Mit Ausnahme weniger Webseiten-Kategorien zeigt sich ein ähnlich hohes Ausmaß von Tracking über Arten von Webseiten hinweg (Abbildung 3). So können zwischen 60 bis 80 Prozent der Zeit im Internet auf Webseiten beobachtet werden, die Wettspiele, Gesundheits- und Ernährungsinformationen oder Immobiliengeschäfte anbieten. Auf Webseiten, die für Karriereplanung, Finanzen oder die Partnersuche verwendet werden, können 40 bis 60 Prozent der Zeit online beobachtet werden. Ausnahmen mit geringeren Prozentsätzen von unter 20 Prozent stellen Webseiten von öffentlichen Institutionen und Regierungen, E-Mail-Anbietern, Messenging-Diensten und Erotikwebseiten dar.

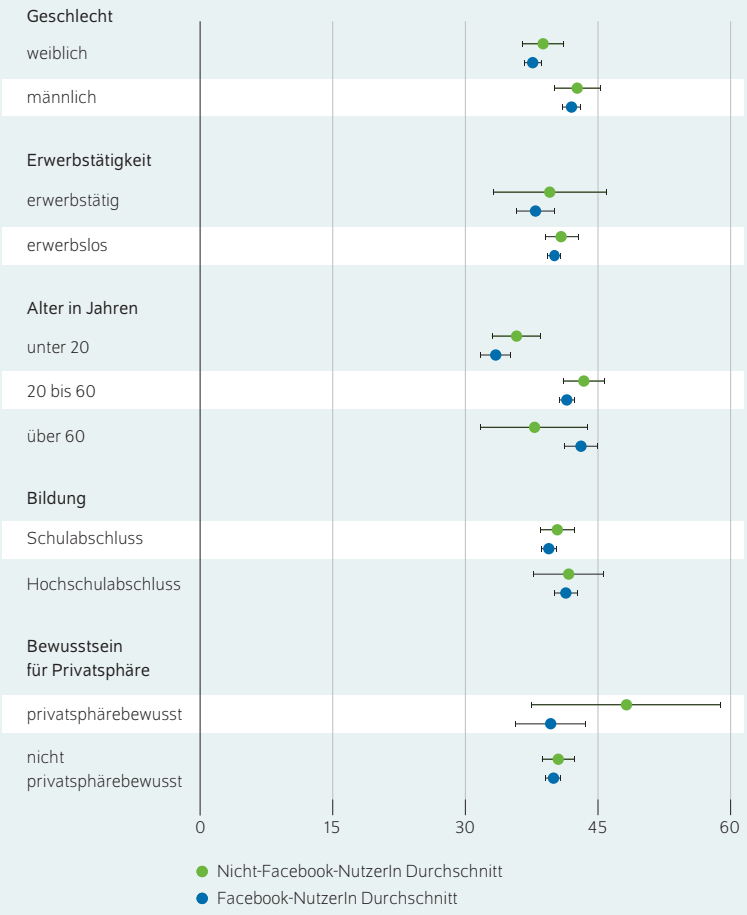
Aus dem Surfverhalten können Plattformen auch über Eigenschaften von Nicht-NutzerInnen lernen

Aufgrund der vielseitigen Inhalte im Internet kann das Gesamtbild des Verhaltens im Internet Aufschluss über Eigenschaften und Vorlieben jedes Einzelnen liefern. Zum Beispiel liegt nahe, dass die Wahrscheinlichkeit, dass ein Kind im Haushalt des Nutzers lebt, mit der Anzahl von besuchten Webseiten, die Kinderprodukte vermarkten, steigt. Werden zusätzlich häufig Internetforen für Väter besucht, steigt die Wahrscheinlichkeit, dass der Nutzer ein Mann, vermutlich mittleren Alters, ist. Aus dem gesamten Nutzungsverhalten, das hunderte Stunden und zahlreiche Webseiten verschiedenster Inhalte umfassen kann, können differenzierte Rückschlüsse auf die persönlichen Nutzereigenschaften gezogen werden.

Um diese Korrelationen sichtbar zu machen, ist es notwendig, die realen Eigenschaften der NutzerInnen zu kennen. Die vorliegende Studie macht sich zu Nutzen, dass die zugrunde liegenden Nutzerdaten demografische Informationen über die teilnehmenden Personen enthalten. So kann quantifiziert werden, wieviel Informationen über individuelle NutzerInnen sich aus deren Nutzerverhalten im Internet gemessen an der Anzahl von Klicks pro Webseite innerhalb eines Jahres herleiten lassen. Hierfür wird

Abbildung 2

Beobachtbares Surfverhalten nach demografischen Eigenschaften und Facebook-Status
Anteil der beobachtbaren Zeit im Internet in Prozent



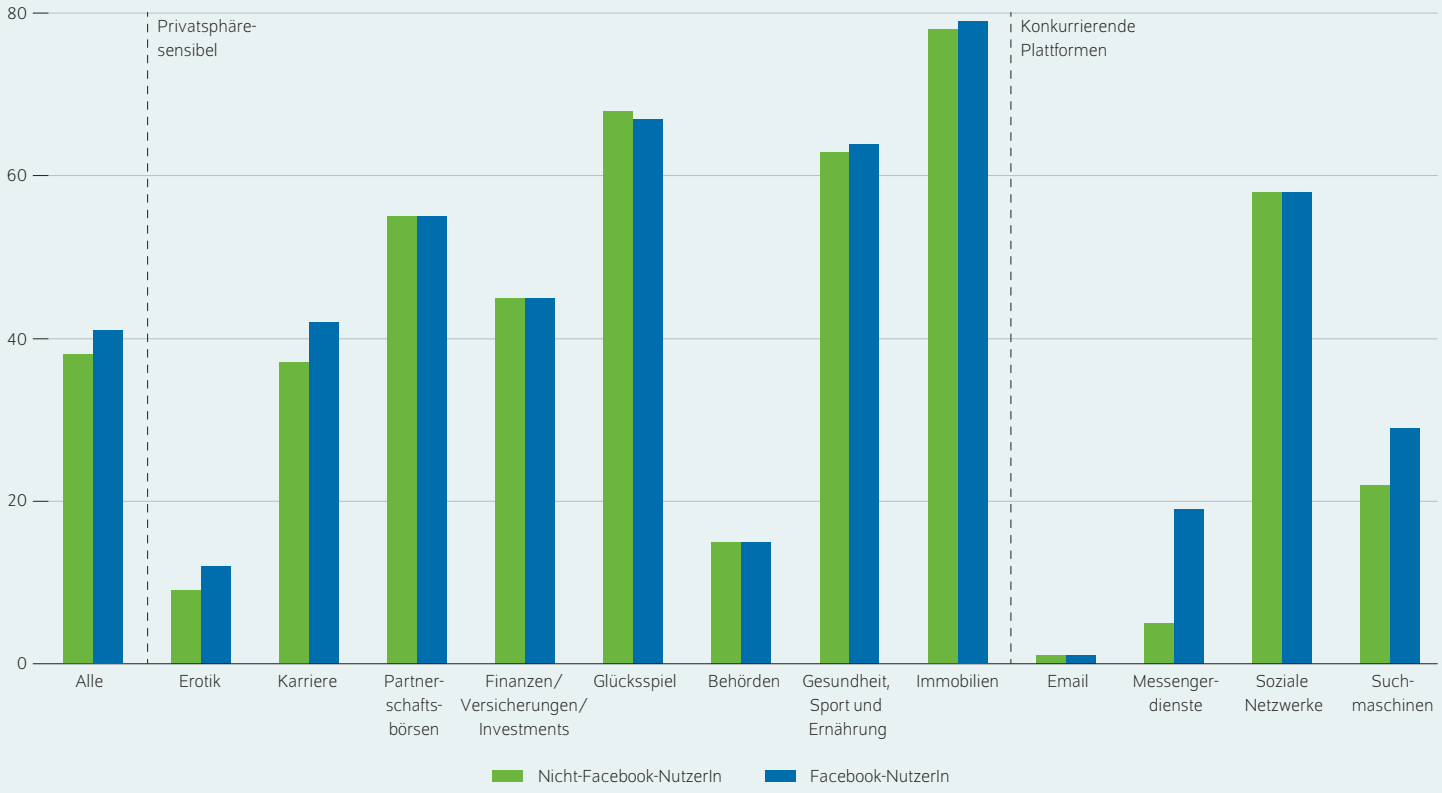
Anmerkungen: Die 95-Prozent-Konfidenzintervalle zeigen die Varianz innerhalb der Personengruppen. Das 95-Prozent-Konfidenzintervall bedeutet, dass in 95 Prozent der Fälle der unbekannte tatsächliche Wert in diesem Intervall liegt. Die Fehlerwahrscheinlichkeit beträgt entsprechend fünf Prozent. Je enger das Intervall, desto genauer ist der geschätzte Effekt.

Quelle: Nielsen Clickstream-Daten, httparchive Tracking-Daten, eigene Berechnungen.

Plattformen wie Facebook können auch das Surfverhalten von Menschen, die nicht dort angemeldet sind, beobachten

Abbildung 3

Tracking-Intensität nach Webseitenkategorie
Anteil der beobachtbaren Zeit im Internet in Prozent



Quelle: Nielsen Clickstream-Daten, httparchive Tracking-Daten, eigene Berechnungen.

Vor allem Aktivität auf Webseiten, die sich mit Immobilien, Glücksspiel und Gesundheit beschäftigen, kann beobachtet werden.

eine Methode des maschinellen Lernens angewandt: das *Extreme Gradient Boosting*, das anhand der Nutzungsdaten die Wahrscheinlichkeit schätzt, dass eine Person zu einer bestimmten Personengruppe gehört.[12]

Die Daten, die Plattformen über ihre NutzerInnen besitzen, beinhalten nicht nur deren Aktivitäten und Vorlieben auf der Plattform. Die meisten NutzerInnen teilen in ihren Profilen persönliche Daten wie ihr Geschlecht, Alter, Wohnort, Ausbildung und mehr. Viele NutzerInnen teilen diese Informationen gerne, da sie so ihre persönlichen Netzwerke über ihr Leben informieren und über soziale Netzwerke in Kontakt bleiben können.

Ein Teil der Menschen entscheidet sich allerdings, diese Plattformen aus Sorge um die eigene Privatsphäre überhaupt nicht zu verwenden. Da sich aber persönliche Eigenschaften anhand von Methoden des maschinellen Lernens und dem beobachteten Nutzerverhalten außerhalb der Plattform schätzen lassen, generieren NutzerInnen sogenannte Datenexternalitäten. Durch das Teilen ihrer eigenen persönlichen Informationen ermöglichen sie also den Plattformen in Verbindung mit ihrem Surfverhalten, Rückschlüsse auf persönliche Eigenschaften von Nicht-NutzerInnen zu ziehen – ohne deren explizites Wissen oder Zustimmung.[13] Hierfür sind drei Voraussetzungen nötig. Zunächst müssen NutzerInnen ihre persönlichen Daten mit der Plattform teilen. Zum Zweiten muss das Nutzerverhalten innerhalb verschiedener Personengruppen von NutzerInnen und Nicht-NutzerInnen ausreichend ähnlich sein. Und schließlich muss die Plattform das Nutzerverhalten per Tracking außerhalb der eigenen Seite beobachten können.

12 Vgl. Trevor Hastie, Robert Tibshirani und Jerome Friedman (2009), a.a.O.

13 Vgl. Mark MacCarthy (2010): New directions in privacy: Disclosure, unfairness and externalities. ISJLP: A journal of law and policy for the information society 6, 425; Jay Pil Choi, Doh-Shin Jeon und Byung-Cheol Kim (2019): Privacy and personal data collection with information externalities. Journal of Public Economics 173, 113–124; Dirk Bergemann, Alessandro Bonatti und Tan Gan (2022): The economics of social data. RAND Journal of Economics, im Erscheinen; Daron Acemoglu et al. (2022): Too Much Data: Prices and Inefficiencies in Data Markets. American Economic Journal: Microeconomics, im Erscheinen.

Um eine mögliche Vorgehensweise für eine Datenextraktion zu simulieren, wird in der Studie ein Algorithmus generiert, der – in der Regel für eine Plattform wie Facebook beobachtbare – persönliche Eigenschaften von NutzerInnen basierend auf deren Surfverhalten schätzt. Dieser Algorithmus wird dann für eine beiseite gelegte Gruppe von NutzerInnen der Plattform Facebook sowie von Nicht-NutzerInnen ausgewertet.

Anhand der vorliegenden Daten kann so die Zugehörigkeit zu verschiedenen Altersgruppen von Facebook-NutzerInnen für 62 bis 80 Prozent der NutzerInnen richtig geschätzt werden, die Anwesenheit von Kindern im Haushalt für 70 Prozent und ein hohes Bildungsniveau für 68 Prozent (Tabelle). Für Nicht-NutzerInnen von Facebook fällt die Korrektklassifikationsrate geringer aus, liegt aber noch bei bis zu 15 Prozentpunkte über der Zufallswahrscheinlichkeit von 50 Prozent. Somit unterscheidet sich zwar die Schätzqualität zwischen NutzerInnen von Facebook, für die die persönlichen Daten ohnehin vorliegen, und Nicht-NutzerInnen von Facebook. Das Teilen von Daten durch Facebook-NutzerInnen verleiht Facebook dennoch die technische Möglichkeit, persönliche Informationen von Personen, die sich dagegen entschieden haben, Facebook zu nutzen, zu schätzen.

Tabelle

Qualität der Schätzung von demografischen Eigenschaften
Qualitätsmaß Korrektklassifikationsrate

		Facebook-NutzerIn	Nicht-Facebook-NutzerIn
Alter	Unter 18	0,80	0,65
	18 bis 24	0,71	0,55
	25 bis 34	0,66	0,56
	35 bis 45	0,62	0,50
	Kinder im Haushalt	0,70	0,57
	Weiblich	0,58	0,53
	Hohes Bildungsniveau	0,68	0,57
	Hohes Einkommen	0,57	0,50
	Erwerbslos	0,55	0,50
	NutzerInnen	3747	1106
	Anzahl Clicks (in 1000)	17671	499

Anmerkungen: Die Korrektklassifikationsrate gibt an, welchen Anteil an allen Beobachtungen ein Algorithmus gewichtet nach der Größe der jeweiligen Personengruppe korrekt klassifiziert. Ein Wert von 0,5 bedeutet dieselbe Qualität wie eine zufällige Klassifikation, ein Wert von 1 bedeutet eine fehlerfreie Klassifikation.

Lesehilfe: Der Algorithmus schätzt mit 65-prozentiger Wahrscheinlichkeit korrekt ein, ob ein Nicht-Facebook-Nutzer unter 18 Jahre alt ist.

Quelle: Nielsen Clickstream-Daten, httparchive Tracking-Daten, eigene Berechnungen

Fazit: Regulierungen versprechen besseren Schutz, auf die Umsetzung kommt es an

Die vorliegende Studie zeigt quantitativ, wie persönliche Informationen auch ohne explizites Wissen vieler InternetnutzerInnen erlangt werden können. So können NutzerInnen technisch auf der Hälfte der besuchten Webseiten und in etwa 40 Prozent der im Internet verbrachten Zeit von Facebook beobachtet werden. Da dies sowohl für NutzerInnen und Nicht-NutzerInnen von Facebook möglich ist, hat Facebook die Möglichkeit, selbst für Nicht-NutzerInnen Profile zu erstellen, die das Schalten zielgerichteter Werbung auch außerhalb von Facebook, zum Beispiel über das Facebook Audience Network, ermöglichen.[14]

Zwar muss seit Einführung der DSGVO die Zustimmung der NutzerInnen zur Nutzung von Cookies eingeholt werden, wie einfach und explizit die Zustimmung vom NutzerInnen eingeholt wird, ist in der Praxis dennoch umstritten.[15] Da die Rechtsdurchsetzung hier insbesondere hinsichtlich Third-Party-Cookies durch die DSGVO effektiver wird, sind neue technologische Entwicklungen bereits in Arbeit. Große Plattformen nutzen ihren Zugang zu NutzerInnen und entwickeln Tracking-Lösungen, die durch von ihnen angebotene Software wie Internetbrowser implementiert werden. Dem verbesserten Datenschutz steht gegenüber, dass bei diesen Lösungen die Datenerhebung und -nutzung noch stärker in den Händen dieser Plattformen liegt.

Andere Initiativen versuchen, Nutzeridentifikatoren zu entwickeln, die mehr Transparenz für NutzerInnen von Webseiten anhand von First-Party-Cookies herstellen könnten. Um das durchaus nützliche Erstellen von Konsumentenprofilen zu ermöglichen, gibt es auch Vorschläge, unabhängige Datentreuhänder zu etablieren, die den Datenschutz und die Nutzung von Konsumentenprofilen in Einklang bringen könnten.[16] Sollten Konsumentenprofile gänzlich verschwinden, ist anzunehmen, dass Plattformen auf stärker kontextbasierte Werbung zurückgreifen. Auch dies kann nützlich sein, zum Beispiel wenn man auf der Suche nach einem bestimmten Produkt ist. Kontextbasierte Personalisierung birgt aber auch die Gefahr, dass kognitive Einschränkungen von NutzerInnen ausgenutzt werden können.[17]

Es bleibt zunächst eine offene Frage, wie effektiv die im Digital Markets Act (DMA) und Digital Services Act (DSA) enthaltenen Regeln nach Ablauf der Umsetzungsfristen ab 2024 durchgesetzt werden. Zum Erkennen und Nachweis von sanktionierbaren Praktiken ist ausreichend qualifiziertes Personal in angemessenem Umfang auf europäischer Ebene essentiell. Schlussendlich wird nach einigen Jahren auch im Rückblick zu evaluieren sein, inwieweit die strikteren Regeln das Angebot und die Qualität von digitalen Inhalten verändern. So hatte die Einführung der DSGVO – nicht zuletzt durch das Festhalten der Plattformen an bestehenden Geschäftsmodellen – auch erhebliche ökonomische

14 Vgl. Informationen auf der Webseite von Meta.

15 Vgl. Competition and Markets Authority (2020), a. a. O.

16 Vgl. Katja Seim et al. (2022): Market design for personal data. Tobin Center for Economic Policy Discussion Paper No. 6.

17 Vgl. Paul Heidhues, Mats Köster und Botond Kőszegi (2021): Steering Fallible Consumers. mimeo.

Kosten.[18] Aus ökonomischer Sicht besteht noch viel Unsicherheit bezüglich der Abwägung zwischen Risiken und Chancen der Datennutzung, die vermutlich auch von den BürgerInnen individuell und unterschiedlich bewertet werden dürften.

18 Vgl. Christian Peukert et al. (2022): Regulatory Spillovers and Data Governance: Evidence from the GDPR. Marketing Science, im Erscheinen; Rebecca Janßen et al. (2022): GDPR and the Lost Generation of Innovative Apps. NBER Working Paper Nr. 30028.

Hannes Ullrich ist wissenschaftlicher Mitarbeiter in der Abteilung Unternehmen und Märkte im DIW Berlin | hullrich@diw.de

Luis Aguiar ist Professor für Management und Digitale Transformation an der Universität Zürich | luis.aguiar@business.uzh.ch

Christian Peukert ist Professor für Digitalisierung, Innovation und geistiges Eigentum an der Universität Lausanne | christian.peukert@unil.ch

Maximilian Schäfer ist Postdoktorand an der Yale University | maximilian.schaefer@yale.edu

JEL: D18, L40, L50, L86, M38

Keywords: Privacy, Tracking, Shadow Profiling, Platforms, Data, Social Media.

IMPRESSUM

DIW Berlin — Deutsches Institut für Wirtschaftsforschung e. V.
Mohrenstraße 58, 10117 Berlin
www.diw.de
Telefon: +49 30 897 89–0 Fax: –200
89. Jahrgang 20. Juli 2022

Vertrieb
DIW Berlin Leserservice, Postfach 74, 77649 Offenburg
leserservice@diw.de
Telefon: +49 1806 14 00 50 25 (20 Cent pro Anruf)

Gestaltung
Roman Wilhelm, Stefanie Reeg, Eva Kretschmer, DIW Berlin

Umschlagmotiv
© imageBROKER / Steffen Diemer

Satz
Satz-Rechen-Zentrum Hartmann + Heenemann GmbH & Co. KG, Berlin

Druck
USE gGmbH, Berlin

ISSN 0012-1304; ISSN 1860-8787 (online)

Abonnieren Sie auch unseren DIW- und/oder Wochenbericht-Newsletter unter **www.diw.de/newsletter**