1440/J
vom 11.07.2018 (XXVI.GP)

Anfrage

der Abgeordneten Stephanie Cox, Kolleginnen und Kollegen

an den Bundesminister für Bildung, Wissenschaft und Forschung

betreffend „Re-Identifizierbarkeit von Personen aus Datensätzen"

BEGRÜNDUNG

Führt die Digitalisierung unserer Gesellschaft zum Verlust der Privatsphäre?

Es gibt zwei wesentliche Treiber der Digitalisierung unserer Gesellschaft:
Erstens das Internet, auf dessen Basis Applikationen wie Google Search, Amazon, Facebook etc. geschaffen wurden, durch die sich immer mehr unserer Aktivitäten von der physischen in die digitale Welt verschieben. (Ein Trend, der durch Smartphones, dank derer wir jederzeit mit dem Internet verbunden sind, noch verstärkt wird.)
Zweitens die Tatsache, dass wir immer mehr und mehr Dinge mit Sensoren (z.B. Kameras) oder „künstlicher Intelligenz" (z.B. „machine vision", „natural language processing") ausstatten und miteinander vernetzen („Internet of Things"), wodurch diese Dinge ihr Umfeld „wahrnehmen" und mit anderen Dingen „kommunizieren" können. Insofern besteht kaum Zweifel daran, dass unsere Gesellschaft in den nächsten Jahren bzw. Jahrzehnten weitgehend digitalisiert wird.
Das Institut für Technikfolgen-Abschätzung der ÖAW sowie das AIT kommen in ihrem Bericht *„Foresight und Technikfolgenabschätzung: Monitoring von Zukunftsthemen für das Österreichische Parlament"* vom Mai 2018 zu demselben Ergebnis: *„Technische Entwicklungen, individuelles und gesellschaftliches Nutzungsverhalten sowie politische Rahmenbedingungen stellen Anonymität im öffentlichen Raum zunehmend in Frage."* Des Weiteren schreiben die Berichtsverfasser_Innen: *„Anonymität stellt aber einen wesentlichen Faktor für freie Meinungsbildung, Entwicklung abweichender Verhaltensweisen und Gedanken als Kern gesellschaftlicher Entwicklung sowie für den Minderheitenschutz und somit für die Demokratie dar. Demokratie ist ohne Anonymität (in ihren unterschiedlichen Facetten von freien Wahlen, Berufsgruppenschutz von JournalistInnen (Kaye 2015), RechtsanwältInnen, PolitikerInnen, DiplomatInnen bis zu SicherheitsexpertInnen usw.) nicht möglich."* (S.17).

Es werden immer mehr Daten erzeugt

Über 4 Milliarden Menschen nützen mit heurigem Jahr das Internet. Google hat in seinem Gründungsjahr (1998) rund 10.000 Suchanfragen pro Tag verarbeitet.[1] Heute sind es rund 68.000 Suchanfragen pro Sekunde oder rund 6 Milliarden Suchanfragen

[1] Battelle, J., (2005) *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*. New York: Portfolio.

pro Tag und jede Sekunde werden rund 74.000 Youtube Videos angesehen.[2] Insofern ist es nicht verwunderlich, dass laut einem Bericht von Cisco (2017)[3] der globale mobile Datenverkehr („global mobile traffic") von 4.4 Exabyte pro Monat im Jahr 2015 auf 7.2 exabyte im Jahr 2016 stieg und in den letzten 5 Jahren um das 18-fache wuchs (1 exabyte sind 10^{18} oder 1,000,000,000,000,000,000 bytes). Der jährliche globale IP Verkehr („annual global IP traffic") wird bis 2021 vermutlich 3,3 Zetabyte erreichen[4] (1 Zetabyte sind 10^{21} oder 1,000,000,000,000,000,000,000 bytes).

Je mehr Daten erzeugt werden, desto gläserner wird der Mensch

Wieso ist der Anstieg von erzeugten Daten nun problematisch für unsere Privatsphäre? Die Antwort liegt auf der Hand: Je mehr Daten erzeugt werden, die eine bestimmte Person betreffen, desto mehr weiß man potentiell über diese Person. Da jeden Tag Unmengen an Daten über Jede/n von uns erzeugt werden,[5] stellt sich die Frage, was man tun kann, damit diese Daten nicht auf einzelne Personen rückführbar sind. Die häufigste Antwort lautet: „Pseudonymisierung".

Reicht Pseudonymisierung aus, um unsere Privatsphäre zu schützen?

Die Regierung hat in diesem Jahr beispielsweise das „Datenschutz-AnpG Wissenschaft und Forschung" beschlossen, wonach – teils sensible – Daten zu bestimmten Zwecken an bestimme Einrichtungen in *pseudonymisierter Form* weitergegeben und verarbeitet werden dürfen. Die Idee hinter der Pseudonymisierung von Daten lässt sich leicht anhand des folgenden Beispiels illustrieren:
Der Name einer Person, z.B. „Max Mustermann", wird ersetzt durch eine scheinbar willkürliche Anordnung von Zeichen, z.B. „MDEyMzQ1Njc4OWFiY2RIZg+CU&g=". (Dieses Kürzel ist in Österreich auch bekannt unter der Bezeichnung „bereichspezifisches Personenkennzeichen".) Werden nur diese Kürzel weitergegeben oder veröffentlicht, erscheint es auf den ersten Blick unmöglich herauszufinden, welche Person hinter welchem Kürzel steckt. Fakt ist allerdings, dass sich jede Art von Pseudonymisierung durch nur wenige, zusätzliche Informationen überwinden lässt. Ein Beispiel:
Damit unsere Smartphones Internetzugang und Empfang haben, brauchen wir Telekommunikationsunternehmen. Diese Unternehmen wissen natürlich nicht, welcher Person welches Smartphone gehört, das sich mit ihren Netzwerken

---

[2] http://www.internetlivestats.com/google-search-statistics (Abgerufen am 14.6.2018).
[3] https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html. (Abgerufen: 14.6.2018)
[4] https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html?referring_site=RE&pos=2&page=https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html. (Abgerufen am 14.6.2018).
[5] Z.B. indem wir etwas auf Facebook „liken", bei Google Suchanfragen stellen, auf Amazon Artikeln kaufen, auf Netflix einen Film ansehen, mit einem „Uber" zu einem Geschäftstermin fahren, mit der Bankomatkarte Mittagessen bezahlen, eGovernment Services in Anspruch nehmen oder ganz allgemein im Internet surfen.

verbindet. (Ähnlich wie im obigen Beispiel werden Konsumenten etwa pseudonymisiert.) Problematisch wird es aber, wenn diese Unternehmen (oder Anbieter_Innen von Apps mit „location tracking") wissen, zu welcher Zeit sich welches Gerät an welchem Ort befindet, denn der Ort eines Geräts um 4 Uhr Früh ist vermutlich das Zuhause der Person und der Ort, an dem das Smartphone sich um 10:30 befindet, ist vermutlich der Arbeitsort. Da es nicht viele Personen gibt, die an derselben Adresse wohnen, arbeiten etc., ist es Wissenschaftler_Innen bereits 2008 gelungen, mit nur zwei solcher „cell info records" – unter den tausenden, die jeden Tag gesammelt werden – die Hälfte aller Telefonnutzer_Innen zu identifizieren (dh. herauszufinden, welche Person hinter welchem Pseudonym steckt). Bereits vier solcher „cell info records" erlaubten die Identifikation von 95% der Telefonnutzer_Innen.[6] Mit dem gleichen Prinzip konnten auch 87% der US Bürger_Innen nur anhand von Geburtsdatum, Geschlecht und Postleitzahl identifiziert werden.[7] In einer anderen Studie[8] wurden drei Monate lang "credit card records" von 1,1 Million Menschen analysiert, wobei nur vier Überweisungen, von denen Zeit und Ort aufgezeichnet wurden („spatiotemporal points"), reichten, um 90% der Personen eindeutig zu identifizieren. Ein Drittel der Netflix User konnten 2015 trotz vermeintlicher Anonymisierung (d.h. trotz Pseudonymisierung) der veröffentlichten Nutzerdaten identifiziert werden. Alles, was es dafür brauchte, war das Wissen um drei bis vier gesehene Filme, die überdies auch Aufschluss über politische Orientierung u.a. der jeweiligen Personen gaben.[9]

Ein zweites Beispiel betrifft unmittelbar Gesundheitsdaten:
Eine Wissenschaftlerin der Carnegie Mellon Universität schaffte es mit geringem Aufwand, die Gesundheitsdaten (z.B. Diagnosen, Medikation, Eingriffe) des damaligen Gouverneurs von Massachusetts – William Weld – herauszufinden. Die „Group Insurance Commission" – zuständig für die Gesundheitsversicherung der staatlicher Angestellten – hat spezifische Gesundheitsdaten gesammelt und diese pseudonymisiert – im Glauben, die Daten seien anonymisiert – an Wissenschaftler_Innen weitergegeben und an andere Unternehmen verkauft. Diese Datensätze enthielten u.a. Postleitzahl, Geburtsdatum und Geschlecht. Die Wissenschaftlerin kauft sich um 20 Dollar die „voter registration list" von Cambridge, Massachusetts. Diese enthielt u.a. ebenfalls Postleitzahl, Geburtsdatum und Geschlecht. Da nur sechs andere Personen in Massachusetts dasselbe Geburtsdatum wie der Gouverneur hatten, nur drei dieser Personen männlich waren und er von diesen drei Personen die einzige mit passender Postleitzahl war, hat bereits die Verbindung von bloß 2 Datensätzen und drei Attributen gereicht, um aus den pseudonymisierten Gesundheitsdaten auf die Krankheitsgeschichte und den Gesundheitszustand des Gouverneurs zu schließen.[10]

---

[6] *Montjoye* (2013), The privacy bounds of human mobility.
7 *Sweeney* (2000), Simple Demographics Often Identify People Uniquely.
8 *De Montjoye, Radaelli, Singh, Pentland* (2014), Unique in the shopping mall: On the reidentifiability of credit card metadata.
9 *Narayanan and Shmatikov* (2015), Robust Deanonymization of Large Sparse Datasets.
[10] *L. Sweeney* (2002), k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems.

Dies erklärt vermutlich auch die Weigerung von Frau Ministerin Hartinger-Klein, die Gesundheitsdaten (bzw. „ELGA-Daten") im Rahmen des neuen FOG in pseudonymisierter Form zu öffnen, und weshalb ÖVP und FPÖ kurzfristig einen Entschließungsantrag einbrachten, nach welchem u.a. die Anonymisierung von Daten für die Verwendung zu wissenschaftlichen Zwecken nötig sein soll. (Echte anonymisierte Daten – iSd der DSGVO – würden überdies auch nicht unter die DSGVO und entsprechende nationale Gesetze fallen, da bei diesen der Datenschutz automatisch gegeben ist.)

Die unterfertigenden Abgeordneten stellen daher folgende

Anfrage

1. Wurden bereits Studien in Auftrag gegeben, die das Problem der Re-Identifizierbarkeit von einzelnen Personen aus pseudonymisierten Datensätzen behandeln?
    a. Falls ja, wurden diese Studien veröffentlicht?
    b. Falls ja, welche Studien waren das und was waren die Ergebnisse?
    c. Falls nein, wieso nicht?
    d. Falls nein, ist geplant, entsprechende Studien in Auftrag zu geben?
        i. Falls nein, wieso nicht?
2. Hat Ihr Ministerium eine Strategie, wie man mit diesem Problem der Re-Identifizierbarkeit von einzelnen Personen aus pseudonymisierten Datensätzen umgehen will?
    a. Falls ja, wie sieht diese Strategie aus?
    b. Falls ja, bis wann soll diese Strategie umgesetzt werden?
    c. Falls nein, wieso nicht?
3. Wird Ihr Ministerium konkrete Maßnahmen setzen, um diesem Problem der Re-Identifizierbarkeit zu begegnen? Bitte um ausführliche und getrennte Beantwortung der folgenden Fragen (a.-c.) für i) den öffentlichen Sektor und ii) die Privatwirtschaft.
    a. Falls ja, welche konkreten Maßnahmen sollen ergriffen werden?
    b. Falls ja, bis wann sollen diese Maßnahmen umgesetzt werden?
    c. Falls nein, wieso nicht?
4. Gibt es bereits Schulungen, Richtlinien oder Checklisten für Mitarbeiter_Innen, die mit der Datenveröffentlichung betraut sind, um diese bei der Einordnung bzw. Kategorisierung der Re-Identifikationsgefahr von Daten nach

Veröffentlichung zu unterstützen? (Eine beispielhafte Checkliste findet sich etwa in Cormode (2015), The confounding problem of private data release. DOI: 10.4230/LIPIcs.ICDT.2015.1)

a. Falls ja, wie sehen diese Schulungen, Richtlinien oder Checklisten aus?

b. Falls ja, welches Ausmaß haben diese Schulungen und welche Mitarbeiter_Innen erhalten diese Schulungen?

c. Falls ja, wie wird sichergestellt, dass Richtlinien oder Checklisten verwendet werden? (Wird die Verwendung z.B. dokumentiert?)

d. Falls nein, wieso nicht?

5. Wird Ihr Ministerium Daten künftig nur noch mit Hilfe von Methoden veröffentlichen, die die echte Anynomisierung von Personen (iSd. DSGVO) – und damit die Nicht-Rückführbarkeit von Daten auf eine Person – sicherstellen (z.B. „k-anonymity protection model“[11] oder vergleichbare Modelle)?

a. Falls ja, welche konkreten Methoden sollen angewendet werden?

b. Falls ja, bis wann soll diese Art der Veröffentlichung von Daten – als allgemeine Regel bzw. Praxis – umgesetzt werden?

c. Falls ja, wie soll sichergestellt werden, dass diese Methoden eingehalten werden (z.B. Dokumentationspflicht, Sanktionierung von Rechtsbrüchen)?

d. Falls nein, wieso nicht?

6. Plant Ihr Ministerium, die Erforschung neuer Innovationen und Methoden zu fördern, die dieses Problem der Re-Identifizierbarkeit lösen könnten? (Das österreichische Start Up „Mostly.ai“ arbeitet z.B. an der Erzeugung „synthetischen Daten“ aus bestehenden Datensätzen, wodurch trotz Anonymisierung eine weitere Verwertung der Daten ermöglicht wird.)

a. Falls ja, welche konkreten Maßnahmen wollen Sie setzen und bis wann?

b. Falls ja, welche Innovationen sollen gefördert werden?

c. Falls ja, in welcher Form soll gefördert werden?

d. Falls nein, wieso nicht?

7. Wird Ihr Ministerium der Regierung ein Gesetz vorschlagen, nach dem nur Methoden der Datenveröffentlichung genützt werden dürfen, die die echte Anynomisierung von Personen (iSd. DSGVO) sicherstellen?

a. Falls ja, was sollen die wesentlichen Inhalte des Gesetzes sein?

---

[11] *L. Sweeney* (2002), k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. [Online: https://epic.org/privacy/reidentification/Sweeney_Article.pdf].

b. Falls ja, soll das Gesetz sowohl den öffentlichen Sektor, als auch die Privatwirtschaft verpflichten?

    i. Falls nein, wieso nicht?

c. Falls ja, soll die Rechtslage für den öffentlichen Sektor und für die Privatwirtschaft unterschiedlich ausgestaltet sein?

    i. Falls ja, inwiefern und wieso?

d. Falls ja, welche konkreten Methoden der Datenveröffentlichung sollen gesetzlich verankert werden?

e. Falls ja, bis wann sollen diese Vorschläge gemacht werden?

f. Falls ja, inwiefern sollen z.B. Dokumentationspflichten eine Rolle im Gesetz spielen, um die Einhaltung der Regelungen sicherzustellen und welche Sanktionen soll es bei Rechtsbruch geben?

g. Falls nein, wieso nicht?

8. Wird Ihr Ministerium ganz allgemein eine Änderung bestehender oder die Erlassung neuer Normen – z.B. Gesetze, Verordnungen – (insb. Datenschutzanpassungsgesetzen) vorschlagen, um das Risiko der Re-Identifizierbarkeit von Personen aus pseudonymisierten Datensätzen zu minimieren?

    a. Falls ja, was soll der wesentliche (neue) Inhalt dieser Normen sein?

    b. Falls ja, sollen diese (neuen) Normen sowohl den öffentlichen Sektor, als auch die Privatwirtschaft verpflichten?

        i. Falls nein, wieso nicht?

    c. Falls ja, soll die Rechtslage für den öffentlichen Sektor und für die Privatwirtschaft unterschiedlich ausgestaltet sein?

        i. Falls ja, inwiefern und wieso?

    d. Falls ja, welche Normen sollen geändert oder neu erlassen werden?

    e. Falls ja, bis wann sollen diese Normen dem Nationalrat per Regierungsvorlage vorgeschlagen werden?

    f. Falls nein, wieso nicht?

9. Im Zusammenhang mit dieser Anfrage fragt sich auch, wie das folgende Ziel im Regierungsprogramm zu verstehen ist: „*Transparenz des Bürgers über jene Daten, die über ihn öffentlich verfügbar sind (im Rahmen von oesterreich.gv.at)*“?

a. Welche Daten über bzw. von BürgerInnen sollen veröffentlicht werden? (Bitte um abschließende Aufzählung aller betroffenen Daten bzw. Datensätze und Attribute.)

b. In welcher Form und mit welchen Methoden sollen Daten über BürgerInnen veröffentlicht werden?

   i. Falls Daten in pseudonymisierter Form veröffentlicht werden sollen, wie stellen Sie sicher, dass BürgerInnen aus diesen Datensätzen nicht re-identifizierbar sind?

   ii. Falls Daten in anonymisierter Form veröffentlicht werden, wie stellen Sie sicher, dass tatsächlich vollständige Anonymität gewährleistet ist? (Bitte insb. auch um Erläuterung der technischen Vorgehensweise.)

REPUBLIK ÖSTERREICH Parlamentsdirektion SIEGEL	**Unterzeichner**	Parlamentsdirektion
	Datum/Zeit-UTC	2018-07-12T11:25:35+02:00
	Prüfinformation	Informationen zur Prüfung des elektronischen Siegels und des Ausdrucks finden Sie unter: https://www.parlament.gv.at/siegel
Hinweis	Dieses Dokument wurde elektronisch besiegelt.	