Hochschule für Politik München
an der Technischen Universität München

Hochschule für Politik München | Richard-Wagner-Str. 1 | 80333 München

Prof. Dr. Simon Hegelich
Political Data Science
Technische Universität München
Hochschule für Politik
Richard-Wagner-Str. 1
80333 München

Email: simon.hegelich@hfp.tum.de
Blog:
https://politicaldatascience.blogspot.de
Twitter: @SimonHegelich

17.06.2020

Stellungnahme zur öffentlichen Anhörung des Ausschusses für Recht und Verbraucherschutz zu dem Thema: „Änderung des Netzwerkdurchsetzungsgesetzes (NetzDG)“

Ein Datenzugang für die Wissenschaft: Erfahrungen aus dem Social Science One Projekt von Facebook

Von Seiten der Wissenschaft wird immer wieder betont, dass der Zugang zu verlässlichen Daten ein Kernproblem bei der Analyse der gesellschaftlichen Wirkung von Social Media darstellt. Im Zuge der Kritik an Facebooks Rolle bei den Präsidentschaftswahlen in den USA hat Mark Zuckerberg versprochen, mehr Daten mit der Wissenschaft zu teilen und Facebook hat mit dem „Social Science One“ (SS1) Projekt versucht, neue Standards für die Kooperation von Plattformunternehmen und Sozialwissenschaften zu setzen. Die folgenden Anmerkungen basieren auf den Erfahrungen des einzigen Teilprojekts in SS1, das an einer deutschen Universität angesiedelt war (https://www.ssrc.org/fellowships/view/social-media-and-democracy-research-grants/grantees/hegelich/). Generell gibt es beim Teilen von Daten mit der Wissenschaft eindeutige Zielkonflikte:
Das Geschäft von Plattformunternehmen besteht in der Verarbeitung und Analyse von Daten. Die Analyse des Nutzerverhaltens erlaubt zum Beispiel das gezielte Schalten von Werbung. Werden zu viele Daten für wissenschaftliche Zwecke nutzbar gemacht, dann kann sich das negativ auf das Geschäft der Unternehmen auswirken und Geschäftsgeheimnisse offenbaren.
Gleichzeitig haben Nutzer*innen ein Recht darauf, dass ihre Daten nicht beliebig weitergegeben werden, sondern das Privacy- und Datenschutzstandards (wie u. a. in der DSGVO kodifiziert) eingehalten werden.
Die Wissenschaft hat zunächst einen unbändigen Hunger nach Daten, da klar ist, dass sich viele Fragen erst ergeben, wenn überhaupt bekannt ist, was empirisch erhoben wird.
Gleichzeitig ist eine wissenschaftliche Analyse der digitalen Revolution für die Allgemeinheit, die Politik aber auch für die Unternehmen vorteilhaft, weil nur auf der Basis von gesichertem

Wissen konsistente Regelungen für diesen Bereich gefunden werden können. In vielen Bereichen teilen Plattformunternehmen daher bereits umfangreiche Datensätze mit der Wissenschaft. Wie eine solche Kooperation funktioniert und welche Probleme sie mit sich bringt, kann am Beispiel des Projekts Social Science One nachvollzogen werden: Die Grundidee des SS1-Projektes war, die oben angerissenen Zielkonflikte dadurch zu lösen, dass zwischen Wissenschaft und Unternehmen eine unabhängige dritte Instanz geschaltet wird (die NGO Social Science One). Diese wird von renommierten Wissenschaftler*innen geleitet , die sich gegenüber Facebook mit NDAs (non-disclosure agreements) verpflichtet haben, Geschäftsgeheimnisse zu wahren und im Gegenzug dafür Einsicht in die Datenbestände zu erhalten. In den unterschiedlichen Projekten im Rahmen von Social Science One sollten sowohl Wege gefunden werden, Daten für die wissenschaftliche Community verfügbar zu machen, als auch die bekannten Zielkonflikte zu moderieren. Leider ist SS1 auf ganzer Linie gescheitert. Auch nach dem offiziellen Ende der bewilligten Projekte stehen immer noch nicht die versprochenen Daten zur Verfügung. Zudem scheinen alle Überlegungen hinsichtlich Privacy vorgeschoben zu sein, um eine wissenschaftliche Analyse und Aufarbeitung des Einflusses von Facebook auf die Wahlen zu verhindern, anstatt sie zu ermöglichen.
Eine Einschätzung zu dem Scheitern von SS1 findet sich in der Zeitschrift Nature (Hegelich 2020a, https://www.nature.com/articles/d41586-020-00828-5). Eine umfassende Analyse der von Facebook zur Verfügung gestellten Daten und der angewandten Methode „Differential Privacy“ habe ich hier veröffentlicht: http://politicaldatascience.blogspot.com/2020/03/the-social-science-one-facebook.html (Hegelich 2020b).
Dennoch lassen sich wichtige Lehren aus diesem Projekt ziehen:
Die Idee einer vermittelnden Instanz scheint essentiell: Einerseits sind Wissenschaftler*innen datenhungrig und nicht notwendigerweise daran interessiert, Unternehmensinteressen oder Datenschutzrechte von Nutzer*innen zu wahren (siehe Cambridge Analytica). Wie das Beispiel SS1 zeigt, braucht es aber auch in Richtung der Unternehmen eine dritte Instanz, da selbst ein mehrfach wiederholtes persönliches Versprechen des CEO keine Garantie ist, dass ein Unternehmen kooperiert. Die Verankerung einer gesetzlichen Kooperationspflicht mit der Wissenschaft und eine mit robustem politischen Mandat ausgestattete Institution, würde solchen Kooperationen das notwendige Gewicht verleihen. Darüber hinaus muss es Aufgabe dieser Institution sein, wissenschaftliche Projekte zu identifizieren, bei denen das gesellschaftliche Interesse so hoch ist, dass ein Austausch von Daten zwischen Unternehmen und Wissenschaft notwendig ist. Gleichzeitig muss die Institution gegenüber der Wissenschaft hohe Standards hinsichtlich ethischer und datenschutzrechtlicher Überprüfung der Projekte durchsetzten können. Welche Daten für eine spezifische Frage gebraucht werden und wie solche Daten zur Verfügung gestellt werden können, bei gleichzeitiger Wahrung der legitimen Geschäftsinteressen und unter Einhaltung von Datenschutz und Urheberrecht, kann nur im Einzelfall ermittelt werden. Eine Vielzahl von technischen Verfahren steht zur Verfügung, um die beschriebenen Zielkonflikte abzumildern:
Daten lassen sich anonymisieren oder pseudonymisieren und dadurch häufig datenschutzkonform machen. Viele wissenschaftliche Analysen brauchen – anders als bei Geschäftsmodellen – nicht die vollständigen Daten: Wenn klar ist, nach welchen Kriterien die Daten zusammengestellt wurden, reicht ein Sample häufig aus. Differential Privacy ist ein recht neuer, vielversprechender Ansatz: Dabei wird verhindert, dass Nutzer*innen anhand ihres Verhaltens oder ihrer Merkmale identifiziert werden können (auch wenn Facebook diesen

Ansatz falsch verwendet hat: Hegelich 2020b). Die Beantwortung vieler Fragestellungen läßt sich zu dem genauso gut mit synthetischen Daten erreichen. In einigen Anwendungsfällen ist es außerdem denkbar, dass überhaupt keine Daten, sondern nur die Ergebnisse von statistischen Berechnungen geteilt werden. Auch eine gestaffelte Kombination von technischen Maßnahmen wäre möglich. Zum Beispiel könnten synthetische Daten allgemein veröffentlicht werden und die interessantesten Analysen dann auf den echten Daten überprüft werden, etc. Welcher Weg in der Praxis zu wählen ist, lässt sich dabei nur am konkreten Fall erörtern.

Wir sehen heute schon, dass gerade wenn es um große Datenmengen geht, die Unternehmen einen nicht einholbaren Wissensvorsprung haben und das in zweierlei Hinsicht: Erstens arbeiten sie täglich mit diesen Daten und zweitens stehen zumindest den großen Plattformunternehmen wesentlich mehr Ressourcen für die Datenanalyse zur Verfügung, als einer normalen Universität. Daraus ergibt sich ein klassisches Principal-Agent-Problem, bei dem es den Unternehmen leicht fällt, den eigentlichen Gedanken der Kooperation zum eigenen Vorteil zu nutzen und darüber ein Ungleichgewicht herzustellen.

Persönlich denke ich, dass sich dieser spezifische Konflikt nicht durch ein Kontrollregime lösen lässt. Die betreffenden Systeme sind so komplex, dass eine wirkungsvolle Kontrolle zu vertretbaren Kosten nicht möglich ist. Als Kybernetiker empfehle ich daher ein System der überwachten Selbstkontrolle: Unternehmen sollen ihre Kooperationen mit der Wissenschaft – wie im Gesetzentwurf vorgesehen – transparent machen. Diese Kooperationen sollten (auch) über eine unabhängige dritte Institution abgewickelt werden, die wiederum ihr Tun transparent offenlegt. Außerdem muss in einem solches Setting über Sanktionen bei einer Kooperationsverweigerung und über Anreize für Kooperationen nachgedacht werden.

Ich bin kein Jurist, aber ich denke, dass die Grundzüge eines solchen Kooperationssystems durchaus im NetzDG festgelegt werden könnten. Die vorgesehene Transparenzpflicht und die Mechanismen der Selbstkontrolle gehen bereits in diese Richtung und könnten durch ein Kooperationsgebot ergänzt werden.

Die Vorstellung, dass sich eine solide Kooperationsstruktur zwischen Unternehmen und Wissenschaft kostenneutral aufbauen lässt, ist allerdings illusorisch: Mit Blick auf den Kontrollaspekt muss festgehalten werden, dass es hier immerhin um Unternehmen geht, die zum Teil Milliarden in den Bereich Datenanalyse investieren. Keine staatliche Organisation (zum Beispiel der Bundesdatenschutzbeauftragte) oder wissenschaftliche Vereinigung (wie die Deutsche Forschungsgemeinschaft) hat derzeit die Ressourcen, um auch nur annähernd auf Augenhöhe mit den großen Plattformunternehmen über Fragen wie Datenqualität und Datenzugang zu verhandeln. Zudem muss klar sein, dass Daten nur dann nützlich und sinnvoll angewandt werden können, wenn in der Wissenschaft die dafür benötigten Strukturen geschaffen werden. Wenn es politisch gewollt ist, dass gerade die Sozial- und Geisteswissenschaften stärker datengeleitet vorgehen, dann braucht es einen massiven Ausbau in der Lehre, in der Forschung und in der Infrastruktur.

Der hier vorgestellte kybernetische Ansatz einer Selbststeuerung mit Unterstützung durch eine unabhängig Organisation und allgemeinen Transparenzpflichten ist daher kostenintensiv. Es scheint aber das einzige Szenario zu sein, das unter den gegebenen Bedingungen funktionieren kann. Vor andere Ansätze, wie beispielsweise der Möglichkeit fallbezogener Regelungen, ist insbesondere aus Anwenderperspektive zu warnen. Der Bundesrat hat beispielsweise das Thema „Social Bots“ als weiteren Punkt zur Prüfung vorgeschlagen. Aus wissenschaftlicher Sicht ist dieses Thema nicht für gesetzliche Regelungen geeignet, weil

aktuell mindestens drei unterschiedliche Definitionen von Social Bots existieren, über die derzeit gestritten wird. Ich habe Social Bots immer als Accounts in den Sozialen Netzwerken definiert, die automatisch kommunizieren und vorgeben echte Menschen zu sein. Wie in meiner Stellungnahme für die Enquetekommission Künstliche Intelligenz ausführlich erläutert, kommt man selbst bei dieser Definition in Erklärungsschwierigkeiten (Hegelich 2020c, http://politicaldatascience.blogspot.com/2020/02/argumente-zu-socialbots-okboomer-kein.html). Daher sind wir in der Forschung dazu übergegangen, von nicht-authentischem Nutzerverhalten zu sprechen und zwar unabhängig vom Grad der Automatisierung. Es existiert aber auch eine Definition von Journalisten und einzelnen Wissenschaftlern, die Social Bots erst als solche sehen, wenn ein Grad an Automatisierung erreicht ist, den wir nur von sehr starken KI-Systemen erwarten können. Und leider definieren nach wie vor viele Wissenschaftler*innen Social Bots einfach als die Accounts, die mit entsprechenden automatisierten Verfahren (die noch dazu offenbar schlecht funktionieren) „entdeckt“ werden. An der Fülle von wissenschaftlichen Publikationen zum Thema kann man gleichzeitig ablesen,dass es gar keinen generellen Mangel an Datenzugängen gibt, da die Plattformen schon recht viele Daten für die Forschung zur Verfügung stellen. Ob in diesem Zusammenhang überhaupt eine Lücke in der Datenbereitstellung besteht, lässt sich allerdings erst anhand spezifischer Forschungsfragen klären.

Zudem geht der potentielle Nutzen der Analyse von Sozialen Netzwerken weit über einzelne Themen hinaus. Die Hoffnung ist, dass sich dieser Datenbestand für eine ganz neue empirisch-basierte Sozialwissenschaft nutzen lässt. Die Betonung besonderer Themen ist in diesem Kontext daher m. E. nach eher kontraproduktiv.

Hegelich 2020a: Facebook needs to share more with researchers. Nature 579, 473 (2020). doi: 10.1038/d41586-020-00828-5.

Hegelich 2020b: The Social Science One Facebook Cooperation: A Systemic Failure. http://politicaldatascience.blogspot.com/2020/03/the-social-science-one-facebook.html

Hegelich 2020c: Erläuterungen zu den Fragen der Enquete Kommission "Künstliche Intelligenz" zum Thema Social Bots. http://politicaldatascience.blogspot.com/2020/02/argumente-zu-socialbots-okboomer-kein.html