Indikatorik und Governance-Ansätze zur Analyse und regulatorischen Gestaltung datenbasierter Märkte in Deutschland

Abschlussbericht des Forschungsprojekts fe 11/19 für das Bundesministerium der Finanzen

Kurzfassung, Politikempfehlungen und Executive Summary

Dezember 2020

Zusammenfassung

Die Entwicklung von Geschäftsmodellen hin zu digitalen Plattformen hat Daten zu einem zentralen Bestandteil für neue Angebote werden lassen. Dies gilt insbesondere für „lernende" Produkte und Dienstleistungen im Bereich Künstlicher Intelligenz (KI) sowie Internet of Things (IoT), die auf möglichst große Datenmengen für sinnvolle Anwendungsszenarien angewiesen sind. Die KI-Strategie der Bundesregierung formuliert als Vorhaben, rechtliche, institutionelle und regulatorische Anpassungsbedarfe angesichts von Big Data und datenreichen Märkten zu diskutieren. Ein wesentlicher Anpassungsbedarf --- darin sind sich Expertenkommissionen u.a. aus der EU, Deutschland, dem Vereinigten Königreich, den Vereinigten Staaten und Australien einig --- besteht in der regulatorischen Behandlung digitaler Plattformen.[1] Es ist zu beobachten, dass Märkte für datenbasierte Plattformgeschäftsmodelle zunehmend von jeweils einem Anbieter dominiert werden --- und dass diese dominanten Anbieter fast immer von nicht-europäischen Großunternehmen kontrolliert werden.

An diesem Punkt setzt das vom Bundesministerum der Finanzen beauftragte vorliegende Forschungsprojekt an, welches die Entwicklung einer geeigneten Indikatorik zur Erkennung und Abgrenzung datenbasierter Märkte sowie, darauf aufbauend, von Ansätzen der Daten-Governance zum Ziel hat. Insbesondere bestand der Auftrag darin, eine Vorgehensweise zur Messung der Datengetriebenheit eines Marktes (also einen Test auf Datengetriebenheit) und der dort vorherrschenden Marktdominanz durch einzelne Anbieter zu entwickeln, diese in einer ausgewählten Branche exemplarisch anzuwenden, sowie Überlegungen hinsichtlich einer geeigneten Daten-Governance-Struktur sowie möglicher regulatorischer Instrumente anzustellen.

Die Hauptergebnisse des Projekts lauten wie folgt:

1. Der entwickelte ökonometrische *Test auf Datengetriebenheit eines Marktes* folgt der Grundfrage: **wie lange benötigt ein Anbieter, der aus Nutzersicht ein perfektes Produkt anbietet[2] aber ohne nutzergenerierte Daten über Präferenzen und Eigenschaften der Nutzer auskommen muss, den Konkurrenten mit dem größten Marktanteil einzuholen?** Lautet die Antwort, weniger als 3-5 Jahre, ist ein Markt nicht (hinreichend) datengetrieben. Lautet die Antwort**, länger als 5 Jahre, dann ist der Markt datengetrieben.** Ohne regulatorische Intervention besteht dann keine Hoffnung auf eine Änderung der Marktstruktur, was negative Auswirkungen auf Innovationsanreize sowohl von potentiellen Markteintretern als auch dem marktführenden Unternehmen hat und durch die große Marktmacht des dominanten Anbieters Raum für vielfältigen Missbrauch bietet, was zu Lasten von Nutzern/Konsumenten geht.

2. Der Test auf Datengetriebenheit wurde exemplarisch anhand des **Marktes für Internet-Suchmaschinen** angewendet. Dort hat ein *Discrete-Choice Experiment* mit 821

[1] Siehe u.a. Kommission Wettbewerbsrecht 4.0 (Schallbruch et al., 2019), EU-Sonderberaterbericht (Crémer et al., 2019), Stigler Committee on Digital Plattforms (2019), UK Digital Competition Expert Panel (2019), ACCC Abschlussbericht (2019).

[2] Das bedeutet, dass der Anbieter in allen Produktmerkmalen, welche die Nutzerzufriedenheit beeinflussen, die nutzerfreundlichste Entscheidung trifft (auch wenn ihn das kurzfristig Umsatz kostet, weil er z.B. auf Einnahmen aus Werbung verzichtet).

Teilnehmern gezeigt, dass sowohl eine Verringerung der *Qualität der Suchergebnisse* als auch eine Erhöhung der *Anzahl der Anzeigen* und des *Personalisierungsgrades* der Suchmaschine eine signifikant negative Auswirkung auf die *Nutzerzufriedenheit* hat. Die negative Bewertung von Personalisierung impliziert eine Präferenz für den Schutz ihrer Privatsphäre. Allerdings ergab sich, dass die Probanden das Merkmal *Qualität* ungefähr *doppelt so stark* wie die beiden anderen Merkmale, *Personalisierungsgrad* und *Werbung,* (jeweils auf einer 5-stufigen Skala) gewichten. Das zeigt die **dominierende Bedeutung der Suchmaschinenqualität im Vergleich zu anderen Produktmerkmalen für Nutzerzufriedenheit (und demnach der Nachfrage**) auf diesem Markt.

3. Darüber hinaus zeigen die Ergebnisse signifikante **Wechselwirkungen des Personalisierungsgrades sowohl mit der Art der Suchanfrage als auch mit dem Grad der Transparenz**. Der negative Effekt des Personalisierungsgrades auf die Nutzerzufriedenheit war bei einer Suchanfrage, die mehr private Informationen über den Nutzer preisgibt, signifikant stärker als bei einer unverfänglichen Suchanfrage --- und signifikant stärker, wenn die Datenschutzinformationen transparent (und nicht versteckt) waren.[3]

4. Durch ein *Experiment mit der Suchmaschine* Cliqz aus München, bei dem die Menge der nutzergenerierten Daten, zu der der Suchalgorithmus Zugang hatte, um die Suchanfrage eines Nutzers zu beantworten, künstlich variiert wurde, wurde deutlich, dass der **Zugang einer kleinen Suchmaschine zu mehr nutzergenerierten Daten ihre Suchqualität stark verbessern würde**. Dies gilt insbesondere für seltene Suchanfragen, unabhängig vom genauen maschinell berechneten Maß für Suchqualität. **Für diese gut 70% aller Suchanfragen** konnte keine Qualitätssättigung durch Zugang zu immer mehr nutzergenerierten Daten festgestellt werden. Diese Ergebnisse bzgl. maschinell berechneter Qualitätsmaße von Suchmaschinen wurden durch menschliche Gutachter der Suchergebnisse qualitativ bestätigt: **Immer mehr nutzergenerierte Daten führen bei seltenen Suchanfragen zu immer höherer Qualität**.

5. Zusammenfassend zeigt der Test auf Datengetriebenheit ein **eindeutiges Ergebnis**: **der Suchmaschinenmarkt ist datengetrieben**. Mit deutlich weniger nutzergenerierten Daten als die führende Suchmaschine ist es auf diesem Markt unmöglich, auch auf mittlere Frist einen Marktanteil zu erreichen, der in die Nähe des Marktführers kommt. Damit ist dieser Markt nicht kompetitiv.

6. Mit Blick auf eine geeignete Governance-Struktur der verpflichteten Datenteilung ergab sich, dass **die bestehenden rechtlichen Mechanismen zur Durchsetzung einer Datenteilungspflicht nach dem EU-Wettbewerbsrecht und der Datenübertragbarkeit nach Datenschutz-Grundverordnung (DSGVO) aus Sicht der vorliegenden Arbeit nicht ausreichen**.

[3] Eine Gruppe Probanden wurden gebeten, bei ihren hypothetischen Suchanfragen davon auszugehen, dass sie nach *Burn-out Symptomen* suchen. Die andere Gruppe (mit einer „unverfänglichen" Suchanfrage) sollte sich vorstellen, nach *Restaurants* zu suchen.

7. In jeder Daten-Governance-Struktur müssen von **Regulierungsbehörden drei wesentliche Aufgaben** übernommen werden: die **Untersuchung** von potentiell datengetriebenen Märkten (d.h. die Durchführung des Tests auf Datengetriebenheit), die **Entscheidung**, ob ein Markt datengetrieben ist und welche genauen Daten von wem mit wem auf welche Weise geteilt werden müssen (d.h. die Bewertung des Testergebnisses), sowie die technische und ökonomisch-rechtliche **Durchsetzung** der Datenteilungspflicht.

8. Aufgrund bestehender EU-Verträge erfordert die Gestaltung der Datenteilungspflicht eine Governance-Struktur, die Elemente einer **ökonomisch effizienten Zentralisierung** mit einer **rechtlich notwendigen Dezentralisierung** der Datenteilung verbindet. Unsere Analysen zeigen **drei probate Governance-Strukturen:**

    1. **Relativ zentralisiert**: Die Untersuchung eines potentiell datengetriebenen Marktes und die Durchsetzung der Datenteilungspflicht werden in einer neuen Europäischen Agentur für den Datenaustausch (*European Data Sharing Agency, EDSA*) zentralisiert, während die Entscheidungsbefugnis bei den nationalen Wettbewerbsbehörden in einem Aufsichtsgremium liegt.

    2. **Dezentralisiert**: Man errichtet ein Kooperationsnetzwerk von nationalen Wettbewerbsbehörden (*Data Sharing Cooperation Network, DSCN*), das durch ein *European Data Sharing Board (EDSB),* dem die Präsidenten aller 27 nationalen EU-Wettbewerbsbehörden angehören, koordiniert wird. Das *DSCN* entscheidet über die Datengetriebenheit eines Marktes. Die für die Untersuchung eines potentiell datengetriebenen Marktes am besten geeignete nationale Wettbewerbsbehörde fungiert als federführende nationale Wettbewerbsbehörde (sogenannte *Lead NCA*), welche die Datenteilungspflicht untersucht und nach der Entscheidung des EDSB in der gesamten EU durchsetzt.

    3. **Gemischt**: Die nationalen Wettbewerbsbehörden werden mit der Untersuchung (*Lead NCA*) und der Entscheidungsfindung (*DSCN*) beauftragt. Die zentralisierte *EDSA* ist für die Durchsetzung der Datenteilungspflicht zuständig.

    Bestehende Durchsetzungsansätze haben bereits die Machbarkeit solcher Regelungen gezeigt. Durch die Einbeziehung von Überlegungen zum Datenschutz und zum Schutz des geistigen Eigentums in die Konzeption selbst, bieten die hier vorgeschlagenen Governance-Strukturen einen konkreten Ansatz für die künftige Regulierung von Daten, der rechtliche und ökonomische Erkenntnisse kombiniert.

Politikempfehlungen im Überblick

1. Auf datengetriebenen Märkten haben Wettbewerber einer dominanten Firma ohne politische Intervention keine Chance, auf mittlere Frist einen Marktanteil zu erreichen, der in die Nähe des Marktanteils des Marktführers kommt. **Wir empfehlen daher die Schaffung neuer gesetzlicher Möglichkeiten zur Regulierung von datengetriebenen Märkten**. Konkret empfehlen wir dort die **Einführung einer Teilungspflicht von nutzergenerierten Daten**.

2. Da der Markt für **Suchmaschinen** datengetrieben ist (s. Ergebnis 5 in der Zusammenfassung), **empfehlen wir auf diesem Markt die Einführung einer Datenteilungspflicht** für nutzergenerierte Daten.

3. Bezüglich der **Ausgestaltung der Datenteilungspflicht** empfehlen wir, unabhängig von einem bestimmten Markt, **folgende Grundsätze**:

    1. Es sollten nur **Rohdaten** geteilt werden müssen, die über die Speicherung der maschinellen Interaktion zwischen Nutzer und Anbieter fast kostenlos vom Anbieter gespeichert werden können und üblicherweise gespeichert werden. Die Analyse dieser Daten obliegt jedem Empfänger selbst. Auf dem Suchmaschinenmarkt entspricht das Suchprotokolldaten (*search logs*).

    2. Auf einem datengetriebenen Markt sollten **alle Firmen mit einem Marktanteil von mindestens 30% zum Teilen** ihrer *nutzergenerierten Daten* **verpflichtet** werden. Somit ergibt sich ein Maximum von drei Anbietern pro Markt, die Daten teilen müssen. Diese Zahl sinkt je weiter der Markt monopolisiert ist.

    3. Auf der Empfängerseite sollte **jede Organisation, die auf dem jeweiligen Markt tätig ist oder die erklären kann, wie sie den Nutzern dieses Marktes mit den Daten dienen würde, Zugang zu den geteilten Daten bekommen**. Dies sollte unabhängig von der Organisationsform der empfangenden Partei gelten, also sowohl für gewinnorientierte, nicht-gewinnorientierte und öffentliche Organisationen.

4. Einerseits zeigt unsere Analyse der verfügbaren Mechanismen des Wettbewerbs- und Datenschutzrechts, dass diese nicht ausreichen, um Monopolisierungstendenzen auf datengetriebenen Märkten zu vermeiden. Andererseits berücksichtigen alle drei vorgeschlagenen Optionen für eine Daten-Governance (s. Zusammenfassung) bereits die durch den Datenschutz und den Schutz geistigen Eigentums gesetzten Grenzen. **Wir empfehlen daher, eine der Optionen, inkl. neu zu schaffender Institutionen und Kommunikationskanäle, zu implementieren**.

5. Bei der Abwägung der Vor- und Nachteile zentralisierter und dezentralisierter Governance sehen wir einen Vorteil bei der „gemischten“ Governance-Struktur: die technische Infrastruktur, die zur *Durchsetzung* der Datenteilungspflicht nötig ist, muss nicht zwischen den nationalen Wettbewerbsbehörden dupliziert werden, da sie auf EU-Ebene innerhalb der *EDSA* stattfindet. Gleichzeitig müssen auf EU-Ebene keine neuen *Untersuchungs-* und *Durchsetzungsbefugnisse* geschaffen werden, da die nationalen Wettbewerbsbehörden eine

federführende nationale Wettbewerbsbehörde auswählen, die am besten in der Lage ist, einen bestimmten Fall zu übernehmen. Die nationalen Wettbewerbsbehörden teilen sich somit gemeinsam die Last für die Nutzung der Ressourcen innerhalb des *DSCN*. Aufgrund dieser Kombination von Merkmalen scheint **die „gemischte" Governance-Struktur optimal** zu sein, weswegen wir **diese Option empfehlen**.

6. Aufgrund von Effizienz-, Datensicherheits- und Datenschutzüberlegungen empfehlen wir, **die nutzergenerierten Daten nicht an Organisationen mit Empfängerrecht weiterzuleiten, sondern sie abgeschirmt in einem Datenpool zusammenzufassen**, der von der technologischen Abteilung der *Lead NCA/EDSA* betrieben wird. Organisationen, die ein Recht auf Zugang zu den geteilten Daten haben, bekommen die **Möglichkeit, ihre ML-Algorithmen im Pool trainieren zu lassen**. Nur die Algorithmen der Empfängerfirmen --- und kein Mensch --- bekommen Zugang zu den Rohdaten, können sie aber nicht aus dem Datenpool herausbringen. Stattdessen können sie nur die gewonnenen Erkenntnisse aus ihren Analysen nach außen tragen, so dass die zahlreichen Anbieter miteinander in echtem Wettbewerb um Nutzer stehen.

Executive Summary

The development of business models of digital platforms has made data a central component for new services. This applies in particular to "learning" products and services in the field of artificial intelligence (AI) and the Internet of Things (IoT), which depend on the largest possible volumes of data for meaningful application scenarios. The AI strategy of the Federal Government of Germany includes a plan in which institutional, regulatory, and cultural adaptation needs are discussed for big data and data-rich markets. One major need for adaptation --- on which expert commissions from the EU, Germany, the United Kingdom, the United States, and Australia, among others, agree --- is the regulatory treatment of digital platforms.[4] It can be observed that markets for data-based platform business models are increasingly dominated by one provider each --- and that these dominant providers are almost always controlled by large non-European companies.

This is the starting point of the present research project commissioned by the German Federal Ministry of Finance, which aims to develop a suitable indicator for the identification and delineation of data-based markets and, based on this, approaches to data governance. In particular, the task was to develop a methodology for measuring the data-driven nature of a market (i.e., a test for data drivenness) and the market dominance of individual providers, to apply this procedure in a selected industry, and to explore a suitable data governance structure and possible regulatory implementation.

The main results of the project are as follows:

1. The developed econometric *test for data-driven markets* follows the basic question: **how long does it take a provider who starts without user-generated data on user preferences and characteristics and hypothetically "does everything right"[5] to catch up with the competitor with the largest market share?** If the answer is **"less than 3-5 years," a market is not (sufficiently) data-driven**. If the answer is "**longer than 5 years," then the market is data driven.** In the latter case, the feedback loop by which having more access to data leads to higher quality, which necessarily increases the market leader's market share, is very strong. Without regulatory intervention, there is then no hope of a change in the market structure. This has a negative impact on the incentives for innovation of both potential market entrants and the market leader. Due to the great market power of the dominant provider, it leaves room for multiple abuses to the detriment of users/consumers.

2. The test for data-driven markets consists of two parts: the assessment of the role of different features in shaping the demand of users and the assessment of the quality feedback loop. To illustrate its use in practice, the test for data drivenness was applied in the **market for internet search engines**. There, *a discrete-choice experiment* with 821 participants showed that both a reduction in the *quality of the search results* and an *increase in the*

[4] See references in footnote 1.

[5] This means that the provider makes the most user-friendly decision regarding all product features that influence user satisfaction (even if it costs her/him revenues in the short term).

number of ads and the *degree of personalization* of the search engine have a significantly negative effect on *user satisfaction*. The negative evaluation of personalization implies a preference for the protection of their privacy. However, we found that respondents rated *quality approximately twice as highly* as the other two characteristics, *personalization level* and *advertising* (each on a 5-level scale). This shows the **dominant importance of search engine quality compared to other product characteristics for user satisfaction (and therefore demand)** in this market.

3. Furthermore, the results show significant **interactions of the degree of personalization with both the type of search query and the degree of transparency**. The negative effect of the degree of personalization on user satisfaction was significantly stronger for a health-related search query than for a harmless search query --- and significantly stronger if the privacy information was transparent (and not hidden).[6]

4. In an *experiment with the search engine* Cliqz from Munich, the amount of user-generated data to which the search algorithm had access to in order to answer a user's search query was artificially varied. It showed that **giving a small search engine access to more user-generated data would greatly improve its search quality**. This is especially true for rare search queries, regardless of the exact measure of search quality. For these **more than 70% of all search queries**, no quality saturation could be determined through access to more and more user-generated data. Human evaluators of the search results qualitatively confirmed these results based on machine-calculated quality measures of search engines: **More user-generated data lead to higher quality for rare search queries**.

5. In summary, the test for data drivenness shows a clear result: **the search engine market is data driven**. With significantly less user-generated data than the leading search engine, it is impossible to achieve a market share on this market that comes close to the market leader, even in the medium term. Therefore, this market is not competitive.

6. With regard to an appropriate governance structure for mandatory data sharing, we found that the **existing legal mechanisms for enforcing a data-sharing obligation under EU competition law and for facilitating data portability under the GDPR are not sufficient**.

7. In any data-governance structure, **regulators must perform three essential tasks**: **investigating** potentially data-driven markets (i.e., performing the test for data-drivenness), **deciding** whether a market is data driven and exactly which data must be shared by whom, with whom, in what way (that is, evaluating the test result), and technically implementing and legally **enforcing** the data sharing obligation.

8. Due to institutional limitations resulting from the EU Treaties, the design of the data-sharing obligation requires a governance structure that combines elements of an **economically efficient centralization** with a **legally necessary decentralization** of data sharing. Our analyses show **three feasible governance structures**:

[6] "Health-related" search query: Search for *burnout symptoms*. "Harmless" query: search for *restaurants*.

1. **Relatively centralized**: The investigation of a potentially data-driven market and the enforcement of the data-sharing obligation will be centralized in a new *European Data Sharing Agency (EDSA)*, while the joint decision-making power of the national competition authorities will lie with a supervisory body.
2. **Decentralized**: A *Data Sharing Cooperation Network (DSCN)* will be established, coordinated by a *European Data Sharing Board*, which will include the presidents of all 27 national competition authorities. The *DSCN* decides on the data-driven nature of a market. The national competition authority best placed to investigate a potentially data-driven market acts as the lead national competition authority (so-called *Lead NCA*), which investigates and enforces the data-sharing obligation throughout the EU.
3. **Mixed**: The national competition authorities are charged with investigation (*Lead NCA*) and decision making (*DSCN*). The centralized *EDSA* is responsible for the enforcement of the data-sharing obligation.

Existing enforcement approaches in data protection and consumer law have already demonstrated the feasibility of such arrangements. By incorporating data protection and intellectual property considerations into the governance design itself, the governance structures proposed here offer a concrete approach to future data regulation that combines legal and economic insights and can be easily taken up by policy makers.

The report leads to the following **policy implications**:

1. In data-driven markets, competitors of a dominant firm have no chance without political intervention to achieve a market share close to that of the market leader in the medium term. Therefore, **we recommend the creation of new legal tools for regulating data-driven markets**. Specifically, we recommend the **introduction of mandatory data sharing of user-generated data**.

2. Because the **market for search engines is data-driven** (see result 5 above), **we recommend the introduction of a data sharing obligation for user-generated data** in this market.

3. Regardless of a specific market, we recommend the following **design principles for mandatory data sharing**:

    1. Only **raw data** should have to be shared, which can be stored almost free of charge by the provider via the automated storage of the interaction between user and provider. The analysis of this data is the responsibility of each recipient. In the search engine market this corresponds to *search log* data.

    2. In a data-driven market, **all providers with a market share of at least 30% should be obliged to share their user-generated data**. This results in a maximum of three providers per market that have to share data. This number decreases the more the market is monopolized.

    3. On the receiving side, **any organization that is active in the respective market or that can explain how it would serve the users of this market with the data should**

be given access to the shared data. This should apply regardless of the organizational form of the receiving party, that is, both to for-profit, non-profit and public organizations.

4. On the one hand, our analysis of the available mechanisms of competition and data protection law shows that these are not sufficient to avoid monopolistic tendencies in data-driven markets. On the other hand, all three proposed options for data governance (see result 8 above) already take into account the limitations imposed by data protection and intellectual property protection. **We therefore recommend implementing one of the three governance options, including newly created institutions and communication channels**.

5. When trading off the pros and cons of centralized and decentralized governance, we see an advantage in the "mixed" governance structure: the technical infrastructure required to enforce the data-sharing obligation does not need to be duplicated between national competition authorities, as this takes place at EU level within the EDSA. At the same time, there is no need to create new investigative and enforcement powers at EU level, as the national competition authorities select a lead national competition authority that is best placed to take over a particular case. The NCAs thus share the burden of using the resources within the DSCN. Due to this combination of features, **we regard the "mixed" governance structure optimal and recommend this option**.

6. For efficiency, data security and privacy considerations, **we recommend that user-generated data is *not* forwarded to organizations with recipient rights, but rather that it is consolidated and shielded in a data pool,** operated by the Lead NCA/EDSA's technology department. Organizations that have a right to access the shared data should be given the opportunity to **have their ML algorithms trained in the pool**. Only the algorithms of the receiving companies --- and no human being --- get access to the raw data, but cannot take it out of the data pool. Instead, they can only transfer the findings from their analyses to the outside world, where a multitude of providers can now compete with each other in a meaningful way.