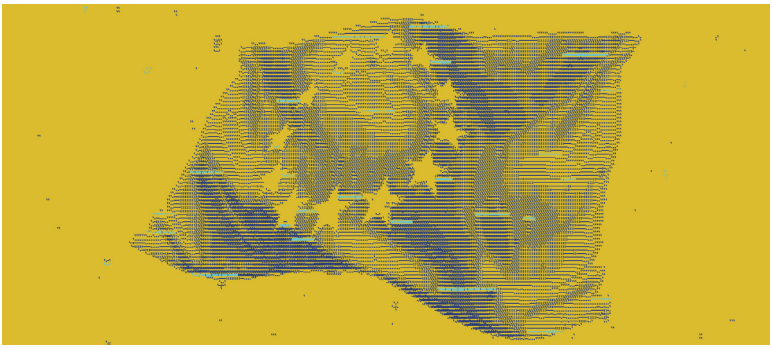
DATA AND DEMOCRACY



The Keys to the Kingdom

Overcoming GDPR concerns to unlock access to platform data for independent researchers

By Mathias Vermeulen

KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY

In October 2020, the Knight First Amendment Institute at Columbia University convened a virtual symposium, titled "Data and Democracy," to investigate how technological advances relating to the collection, analysis, and manipulation of data are affecting democratic processes, and how the law must adapt to ensure the conditions for self-government. This symposium was organized by the Institute's 2019-2020 Senior Visiting Research Scholar, Yale Law Professor Amy Kapczynski, and co-sponsored by the Law and Political Economy Project at Yale Law School.

The essays in this series were originally presented and discussed at this two-day event. Written by scholars and experts in law, computer science, information studies, political science, and other disciplines, the essays focus on three areas that are both central to democratic governance and directly affected by advancing technologies and ever-increasing data collection: 1) public opinion formation and access to information; 2) the formation and exercise of public power; and 3) the political economy of data.

The symposium was conceptualized by Knight Institute staff, including Jameel Jaffer, Executive Director; Katy Glenn Bass, Research Director; Amy Kapczynski, Senior Visiting Research Scholar; Alex Abdo, Litigation Director; and Larry Siems, Chief of Staff. The essay series was edited by Glenn Bass with additional support from Lorraine Kenny, Communications Director; A. Adam Glenn, Writer/Editor; and Madeline Wood, Communications and Research Coordinator.

The full series is available at knightcolumbia.org/research/

"Although we have more data than ever before, a smaller percentage of that data is available to researchers than ever before."[1]

INTRODUCTION

ACADEMICS, JOURNALISTS, civil society organizations, and regulators have increasingly been trying to answer a range of different questions about the role of online platforms in enabling, facilitating, and amplifying the spreading of disinformation online. To what extent are states and nonstate actors actually trying to manipulate elections or democratic debates online? What is the exact role of the infrastructure of some of these platforms—such as ad-targeting mechanisms or content recommendation algorithms—in amplifying and disseminating disinformation? And to what extent are any of the voluntary measures that platforms are implementing to counter disinformation actually effective?

Access to platform data is seen by many of these stakeholders as a necessary condition to not only start answering some of these questions but also to—where relevant—formulate policy interventions that would hold platforms accountable if and when they are responsible for facilitating or exacerbating harms caused by disinformation. Importantly, access to data has been highlighted as a key issue by European politicians and regulators about platform governance in general.

However, in general, platforms have been reluctant to provide such data

voluntarily. Platforms have at times invoked Europe's General Data Protection Regulation (GDPR) as a key obstacle that prevents them from sharing platform data, including personal data, with independent researchers. In general, their argument rests on the assumption that the GDPR lacks clarity regarding whether and how companies might share data with independent researchers, which results in a conservative attitude to making such data available, given the penalties that can be incurred for violating the GDPR. Additionally, the release of such data could result in evidence-based criticism, public outcry, and regulatory action, which means there is currently little to no incentive for platforms to hand over such data.

This article argues that there is a pressing societal need for a legally binding mechanism that provides independent researchers with access to a range of different types of platform data. Such a mechanism would tilt the incentive structure toward more disclosure and would correct the enormous information asymmetry between platforms and third parties. It would also present a first step toward an auditing regime for platforms and would provide more legal certainty for platforms to hand over personal data to researchers in compliance with Europe's privacy laws.

The first draft of this paper was published in November 2020[2] and recommended that the EU create a legal obligation for platforms to make platform data available to independent researchers in order to achieve this goal. Since the publication of that draft paper, the European Commission published its proposal for a Digital Services Act (DSA),[3] which included such a legal provision. Article 31 of the DSA would oblige specific categories of platforms to make data available to independent researchers who act on behalf of a regulator. Such access would be required to monitor and assess the compliance of platforms with new obligations that the EU intends to impose on those platforms in the DSA.

Section I of this article will outline why access to a wide range of platform data for scientific researchers is vital to analyze and mitigate the potential harms resulting from disinformation. Section II will highlight the limits of self-regulation in this context and describe why voluntary mechanisms created by platforms to provide such data have not been sufficient. Section III will highlight how the GDPR currently would allow platforms to share personal data based on consent or legitimate interest, while Section IV argues

that the creation of a separate compelled-data-access regime as proposed in the DSA would add a crucial legal framework to secure more legal certainty for both platforms and researchers. However, this framework for compelling platforms to make data available does not specify what the technical, logistical, and legal parameters are for ensuring that such data access protects user privacy and adheres to best practices for ethical and responsible research. A code of conduct on GDPR-compliant access could solve a number of concerns in this area and prove useful as a model outside the EU as well.

I. ACCESS TO DATA FOR SCIENTIFIC RESEARCHERS IS VITAL TO ANALYZE AND MITIGATE THE HARMS RESULTING FROM DISINFORMATION

OVER THE PAST COUPLE of years, independent researchers whose work aims to analyze, detect, expose, or prevent the creation, facilitation, dissemination, and impact of disinformation online have systematically requested access to a wide variety of platform data in order to move research efforts from primarily descriptive research to causal research.[4] A broad set of actors could benefit substantially from increased access. Journalists, fact-checkers, NGOs, or digital forensics experts perform vital watchdog functions in our democracies, and their ability to improve the public understanding of the challenges of—and potential solutions to—disinformation would be significantly enhanced as a result.[5] In Europe, however, the GDPR carves out specific derogations for "scientific research purposes,"[6] which especially put academic researchers in a privileged position to receive access to platform data.

Different types of platform data

Platform data—or just "data"—is often used as an amorphous all-encompassing term that hides at least three different categories of data, which in the EU would be afforded different levels of legal protection under the GDPR.

The first category consists of predominantly nonpersonal data, which is often proprietary; this can consist of numerical data, metrics, classifiers, or

other types of data that relate to the technical functionalities, design choices, or policy decisions of a particular platform. These can include, for instance, statistics about A/B testing;[7] information about microtargeting options used by advertisers;[8] engagement and impression data of watched videos; data that enables network analysis of a video recommendation system;[9] or internal categorizations of specific pieces of content or metrics regarding the enforcement of specific violations of terms of service. This category also includes data about the parameters that determine the ranking, sequencing, rating, or review mechanisms of content, visual highlights, or other saliency tools.[10] In the EU, platforms are already required to outline these parameters "in order to improve predictability" for business users (such as restaurants, travel agents, or shops) to allow them to "better understand the functioning of the ranking mechanism and to enable them to compare the ranking practices of various providers."[11] The GDPR does not apply to this data insofar as it is not "personal" data.[12]

A second category of data consists of personal data, which according to the GDPR refers to any information relating to an identified or identifiable natural person ('data subject').[13] Essentially, personal data is any type of data that relates to a natural person or can reveal the identity of a user of a specific platform, for example, the account information or IP addresses of individual users. The sharing of such data is subject to a number of rules and safeguards.

A third category of data consists of "sensitive" or "special category" personal data, which according to the GDPR refers to the processing of data that reveals someone's "racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership."[14] It also includes health data or data "concerning a natural person's sex life or sexual orientation."[15] The processing or sharing of such special category data is in principle prohibited, unless an exception to this general principle applies.

Researchers need access to these different categories of data for the purpose of conducting scientific research that can contribute to a number of goals, including, but not limited to (i) the identification of organized disinformation campaigns by state actors; (ii) the identification and understanding of how platform users engage with disinformation and what societal risks this poses, (iii) the identification and understanding of how platforms' own

policies and products can contribute to societal risks posed by disinformation; and (iv) the testing of the effectiveness of voluntary interventions by platforms to counter disinformation on their platforms.

Identifying organized disinformation campaigns

The field of disinformation studies bridges the fields of sociology, media studies, and political science, with more technical domains of data science and cybersecurity. This field "crystalizes around a shared goal of revealing how malicious actors utilize information and communication technologies."[16] One very specific purpose for which access to data is needed by some experts in this field is to attribute organized disinformation campaigns to specific, well-funded, state or nonstate actors. Data that allows for the attribution of information operations is crucial to formulate proportionate (counter) responses that might deter similar behavior from taking place in the future.

Research that focuses on attribution is essentially an extended field of research for cybersecurity scholars and practitioners, who have been very critical about the lack of data from platforms.[17] The detection of manipulative actors who engage knowingly and covertly in what Camille François dubbed "viral deception campaigns" relies on a "cat-and-mouse game of a) identifying threat actors willing and able to covertly manipulate public discourse and b) keeping those actors from leveraging social media to do so, as they refine their strategies to evade detection."[18] François, in her briefing for the U.S. House of Representatives Committee on Science, Space, and Technology, highlighted the "dramatic asymmetry of information" between the platforms on whose infrastructure these influence operations play out, and the rest of the world. There she lambasted some platforms' community standards or terms of service, which she said either "indirectly prevent the type of external research that may lead to detecting and exposing distortive behaviors (e.g., when existing and important safeguards also prevent researchers from collecting the data they'd need to analyze distortive behaviors) or directly seek to prevent it (e.g., with rules explicitly preventing the use of data in order to perform detection of deceptive behavior)."[19]

François received access to nonpublic platform data on behalf of the U.S. Senate Intelligence Committee in order to scrutinize the Russian campaign

targeting the American public around the 2016 presidential election. Her conclusion after seven months of intense research, even with that unprecedented level of access, was that "[t]here remain critical data blind spots … which undermine our preparation for the threats ahead."[20] James Pamment points out that the security teams at the major platforms already have an "open, trusted channel for sharing intelligence on disinformation leads and threat actor tactics, techniques, and procedures"—in contrast to other business areas, which indicates in his view at least the "feasibility of collaborating on a shared repository of analytics and campaign-wide data for … the research community."[21]

Identifying which platform users are engaging with disinformation and why

A recent paper in the Harvard Kennedy School Misinformation Review describes research that could hypothetically be conducted if platform data were more readily available.[22] Researchers describe that they are currently unable to scrutinize the "true reach" of disinformation that is being shared because impression data is not made available.[23] Access to personal data would enable researchers to scrutinize "motivations for sharing misinformation" and enable analyses of "the extent to which people believe misinformation as they share or receive it, as mediated by user characteristics and the sharer-receiver relationship."[24] Access to special category data could enable researchers to understand "to what degree does misinformation play on emotions stemming from ideological, political, racial, or religious biases."[25]

Access to such data could also assess the potential societal risks that are the result of exposure to misinformation, for instance, whether such exposure is "negatively associated with less participation in public discussion" or "less involvement in civic and political activities."[26] In announcing one of Facebook's latest data-sharing efforts, Nick Clegg, vice president for global affairs and communications, and Chaya Nayak, head of Facebook's general research and transparency team, admitted that there is a need for "more objective, dispassionate, empirically grounded research." According to Clegg and Nayak, "We need to better understand whether social media makes us more polarized as a society, or if it largely reflects the divisions that already exist; if it helps people to become better informed about politics, or

less; or if it affects people's attitudes towards government and democracy, including whether and how they vote."[27]

Identifying how platforms enable, facilitate, or amplify disinformation

Platforms have taken a number of measures in the past five years to counter various symptoms of the disinformation problem, including by updating their terms of service to outline which specific activities are prohibited on their platforms. However, most of these interventions focus on the behavior of "bad" external actors; they rarely address the core enabling infrastructure of the platforms that can facilitate the spreading and amplification of disinformation.

This core architecture consists of a series of built-in optimization rules, engagement algorithms, and incentives structures that are intended features of platforms like Facebook or YouTube. Active design choices are made about this core infrastructure with the primary goal of increasing our engagement with content on the site, which increases the delivery of targeted ads.[28] These features determine access to information, its amplification, and ultimately the reach of specific pieces of content. These are the same features that are used—and gamed—by every influencer or online marketer that wants to push a specific message. "Bad actors" in that sense are often using the same tools that legitimate business users have at their disposal.[29]

This is illustrated by the fact that despite all the voluntary actions to counter COVID-19-related misinformation by the platforms, that misinformation is still rampant—perhaps because content promotion and amplification systems are largely left untouched by these actions.[30] According to a recent study published in BMJ Global Health, over 25 percent of the most viewed YouTube videos about coronavirus contain false or misleading information, which has led to suggestions that COVID-19-related misinformation is being pushed more frequently than verified health content. Another investigation by the Institute for Strategic Dialogue (ISD) and the BBC in 2020 found that websites known to host disinformation about COVID-19 had received more than 80 million interactions on public Facebook pages since the start of the year. As a benchmark, in the same period, links to the Centers for Disease Control and Prevention and World Health Organization websites gathered

around 12 million interactions combined.[31] In another area, The New York Times suggested that Facebook's recommendations systems, designed to prioritize the growth of closed groups, "most likely supercharged the QAnon community—exposing scores of people to the conspiracy theory."[32]

When evidence emerges about the enabling role of this core infrastructure in facilitating the amplification and dissemination of disinformation, no action by the platform ensues—unless undeniable evidence is made public. When an internal report at Facebook, obtained by The Wall Street Journal, found that 64 percent of people who joined an extremist group on Facebook only did so because the company's algorithm recommended it to them—no follow-up action ensued.[33] But when The Markup published a story about how Facebook allowed advertisers to target people who were interested in "pseudoscience," the company removed this targeting option.[34] Hence researchers want access to any relevant platform data, which includes proprietary data but also personal and sensitive data, in order to properly scrutinize the role of this core architecture in amplifying and disseminating disinformation.

Testing the effectiveness of voluntary interventions by platforms to counter disinformation

Another crucial reason disinformation scholars want access to data is to assess the effectiveness and proportionality of specific, voluntary interventions by the platforms to counter disinformation. Facebook's fact-checking partners, for instance, "don't know how well their efforts perform at reducing the spread of misinformation."[35] As Gordon Pennycook and David Rand argue: "Experimental investigation of interventions, rather than implementation based on intuitive appeal, is essential for effectively meeting the misinformation challenge."[36] Specific interventions that researchers would want to investigate include labeling news headlines with fact-checking warnings and prompts that nudge users to consider accuracy before sharing. Fact-checkers earlier complained that they don't know the effect of platform strategies, such as reducing the reach of a debunked piece of content or automatically displaying related articles, in dissuading people from sharing a link that contains debunked content.[37] As Michelle Amazeen and Chris Vargo conclude: "Researchers need visibility into these actions to

assess how political ideology, media use, and media literacy interact with the steps platforms are taking to correct misinformation."[38]

A report by the ISD on the activities of social media platforms to counter COVID-19-related misinformation illustrates how platforms have responded quickly to the challenges posed by bad actors, including by setting up information hubs that share verified updates from trusted sources such as the WHO; labeling, downranking and/or removing content flagged as false or misleading by experts; and prohibiting ads that aim to profiteer off the pandemic. Unfortunately, there is no way to independently assess the effectiveness of these interventions. Statistics given by platforms need to be taken at face value, but claimed successes cannot be independently verified. As the ISD notes:

> Without better access to data and insight on companies' decision-making systems, both human- and machine-led, we cannot determine with certainty why some areas of policy appear more effective or better enforced than others. ... Without such data any conclusions drawn about the response of these platforms must rely on some element of extrapolation and inference.[39]

Voluntary mechanisms created by the platforms to provide access to data are not sufficient

Platforms have always made certain types of data available to a limited extent to the research community via public APIs, or application programming interfaces. These APIs have been regarded as "perhaps the most valuable resource that platforms have offered to third-party researchers."[40] They allow third parties to request machine-readable data in bulk on a range of relevant topics. Twitter, for instance, has for years provided relatively generous access to public tweets in real-time via its open streaming API and to historical data via its search API, which explains why many studies in the field of disinformation disproportionately focus on the use of that platform.[41] YouTube has a similar access regime, which allows the downloading of public YouTube data about channels, videos, and searches. In the wake of the Cambridge Analytica scandal, Facebook shut down Instagram's public API and substantially reduced the functionalities of Facebook APIs that provided data on public activities on its public events, groups, and pages.[42] Alexander

Sängerlaub, for instance, argued that after this reduction he was unable to replicate his own research on the 2017 German federal elections.[43] From October 2019, Facebook also made data available to pre-vetted researchers via its CrowdTangle API. The data available to researchers through CrowdTangle describes aggregated interactions with Facebook and Instagram posts from public pages, public groups, or public people, including engagement data such as the number of user reactions, shares, comments, and comparisons to a benchmark that illustrates overperformance or underperformance for each kind of interaction.[44]

These APIs all provide access to public data sets, which pose limited GDPR concerns. As pointed out above, however, researchers would require access to more granular data to investigate a wide range of topics. Interestingly, from 2018 onwards, these requests started to resonate among political stakeholders in the EU. In general, platforms' voluntary commitments to make data available to researchers have been characterized by legal and technical snafus. Unilateral changes to earlier APIs with little to no notice to independent researchers have seriously hindered research that would have a clear public interest.[45] Despite the goodwill and genuine efforts from a number of platform representatives and the research community to make data access mechanisms a success, the results have been patchy at best and illustrative of the major power asymmetries between the platforms and the independent researchers.

II. THE LIMITS OF SELF-REGULATION: THE EU'S CODE OF PRACTICE ON DISINFORMATION AND FACEBOOK'S POLITICAL ADS ARCHIVE

The EU's interest in countering disinformation has for a long time focused almost exclusively on countering narratives and false information originating from Russia. This started to change in official documents from 2018 onwards, when the EU announced the creation of a "code of practice on disinformation that would commit online platforms and the advertising industry—among other things—to provide academia with

"access to platform data (notably via application programming interfaces), while respecting user privacy, trade secrets, and intellectual property."[46] This would enable researchers "to better understand the functioning of related algorithms and better analyze and monitor disinformation dynamics and their impact on society."[47]

However, this broad intention did not survive the drafting process of the code. In the final Code of Practice,[48] the signatories "acknowledged" the importance to "take the necessary measures to enable privacy-compliant access to data for fact-checking and research activities" and to "cooperate by providing relevant data on the functioning of their services, including data for independent investigation by academic researchers and general information on algorithms." Hence relevant signatories committed to "support good faith independent efforts to track disinformation and understand its impact" and stated explicitly that "this *will include* sharing privacy protected datasets, undertaking joint research, or otherwise partnering with academics and civil society organizations *if relevant and possible*."[49] Relevant signatories also committed not to "prohibit or discourage good faith research into disinformation and political advertising on their platforms."[50] The signatories only committed to undertake those actions that correspond to the product or service they offer and "their technical capabilities."[51]

In practice, these promises boiled down to providing access to nonpersonal data via APIs to a limited data set, in this case, data about political ads that was made available via an ad archive. From the three platforms that set up political ad archives (Google, Twitter, and Facebook), Facebook's example acts as a cautionary tale about the limits of self-regulation in this area.

According to Facebook, the ad library lets people "easily see how many political and issues ads were run in a given country—as well as aggregated advertiser spend and top searched keywords in the Ad Library."[52] However, while ordinary users could indeed relatively quickly look up individual ads to a certain extent, research access to the library's data was more problematic.

Seventy-five independent researchers, brought together by the Mozilla Foundation, argued that many meaningful data points were missing from the archive, including targeting criteria, impression data, engagement data, and data about microtargeting.[53] Independent researchers and civil society organizations accused Facebook of "blocking the ability of independent

researchers to effectively study how political disinformation flows across its ad platform." In response, and in private, Facebook has "responded to concerns raised about its ad API's limits by saying it cannot provide researchers with more fulsome data about ads—including the targeting criteria for ads—because doing so would violate its commitments under the EU's [GDPR] framework."[54]

Moreover, the tool Facebook built was so flawed that it was seen as "effectively useless as a way to track political messaging."[55] Researchers that originally set out to track political advertising ahead of the European elections in May 2019 "instead ended up documenting problems with Facebook's library after managing to download the information they needed on only two days in a six-week span because of bugs and technical issues, all of which they reported to Facebook."[56] A study by the French government confirmed the same findings.[57] Crucially, the ad library was missing information about the targeting data that determines who sees the ads. When researchers at New York University built a browser extension that collected both the content of the political ads and the targeting data, Facebook eventually sent a cease-and-desist letter to the researchers. The letter ordered them not only to shut down the browser extension but also to delete all the data they had collected.[58]

It is important to dwell on this topic since the ad library and its API have been touted many times by senior executives as an example of how Facebook was a responsible corporate citizen that could be trusted to fix its own problems. In response to Twitter's announcement that it would cease all political ads on its platforms, Facebook's Chief Operating Officer Sheryl Sandberg responded that Facebook does not have to cease political advertising because the platform is "focused and leading on transparency," explicitly citing Facebook's ad archive efforts.[59] By leaving it up to the platforms to decide how they should implement their obligations under the code of practice, the access to data regime for researchers remained an empty box.

By January 2019, the European Commission itself started to admit that it was "deeply concerned by the platforms' failure to provide specific benchmarks to measure progress, by the lack of detail on the actual results of the measures already taken and lack of detail showing that new policies and tools are deployed timely and with sufficient resources across all EU Member

States."[60] In March 2019, European Commissioner Julian King "urged the platforms to do more to improve independent scrutiny" and to ensure that the platforms are "not just marking their own homework."[61]

In its final assessment of the code of practice, the commission singled out the lack of access to data as "a fundamental shortcoming" of the code.[62] The provision of data and search tools required to detect and analyze disinformation cases was seen as "episodic and arbitrary," and did "not respond to the full range of research needs." The voluntary nature of existing data-sharing mechanisms was seen as insufficient and hence "a more structured model for cooperation between platforms and the research community should be developed."[63]

The limits of data access partnerships

To its credit, Facebook has been experimenting with more bespoke and targeted schemes for scholarly data access. The most prominent one is Social Science One, which was announced in April 2018 as a new partnership between independent researchers and Facebook "to study the impact of social media on democracy and elections, generate insights to inform policy at the intersection of media, technology, and democracy, and advance new avenues for future research."[64] This "new approach for industry-academic partnerships" has been pioneered by Gary King (Harvard) and Nate Persily (Stanford), who decided to set up a trusted "third party"—Social Science One—whose members consist of a commission of senior academics who came to an agreement with Facebook on the scope of a research project.

Social Science One's press release in July 2018 initially promised the release of "about a petabyte of data with almost all public URLs Facebook users globally have clicked on, when, and by what types of people, including many links judged to be intentionally false news stories by third party fact checkers."[65] Ultimately, in September 2019, it released a limited data set of 7 gigabytes, comprising 32 million "misinformation" URLs that have been shared on Facebook. The long delay resulted in significant criticism from the original funders of Social Science One, who expressed frustration that the 83 independent scholars whose proposals were selected for funding received "access to only a portion of what they were told they could expect, and this has made it difficult or, in some cases, impossible for them to complete the

approved research."[66] A year and a half after the initial announcement of Social Science One, in December 2019, its co-chairs and its European Advisory Committee stated that "Facebook has still not provided academics with anything approaching adequate data access." They argued:

> The current situation is untenable. Heated public and political discussions are waged over the role and responsibilities of platforms in today's societies, and yet researchers cannot make fully informed contributions to these discussions. We are mostly left in the dark, lacking appropriate data to assess potential risks and benefits. This is not an acceptable situation for scientific knowledge. It is not an acceptable situation for our societies.[67]

The reasons for the delay were twofold. Initially, Social Science One thought that researchers could do their research using Facebook's systems, but according to Persily, its co-chair, "The company did not have structures that could be readily adapted to give parties access to specific data."[68] Hence a customized "unprecedented research infrastructure" had to be built from scratch in order to secure the privacy of the data sets,[69] using differential privacy techniques.

The key rationale put forward by Facebook, however, was that they were constrained by restrictions in both the GDPR and the "consent decree" it operates under with the U.S. Federal Trade Commission.[70] Facebook argued that those restrictions "prevent researchers from analyzing individual level data, even if de-identified or aggregated."[71] Despite these being the key rationales to prevent access to such data, no detailed reasons were provided at that time, which led the Social Science One co-chairs and its European advisory committee to urge the "major digital platforms" to offer formal, written analyses of any legal barriers they claim prevent them from providing access for academic research, including with regards to the European Union's [GDPR]."[72] The academics argued that the GDPR supported "a more permissive interpretation with respect to academic data sharing for public good."[73]

The end result was a compromise, where Social Science One complied "with Facebook's interpretation of the applicable privacy laws,"[74] which resulted in applying differential privacy techniques to the URL data set. This introduced statistical noise and censoring into the data set, which meant that

the usefulness of the research was significantly curtailed.[75] Despite the good faith efforts of many people involved in this project, who genuinely wanted to make it a success, the history of Social Science One is an example of the limits of voluntary data-sharing approaches between platforms and researchers, where access to data is determined exclusively by the interpretation of the GDPR by one dominant platform.

Most recently, Facebook introduced a U.S. 2020 election research initiative,[76] which pairs a group of outside scholars with internal Facebook employees to study Facebook's and Instagram's impact on four key outcomes that have dominated public and academic attention: political participation, political polarization, knowledge and misperceptions, and trust in U.S. democratic institutions.[77] Only Facebook employees will be able to "touch" and analyze the raw data,[78] but "both teams will work together to devise appropriate monitoring systems for assuring the scientific integrity of the research." This is an important, and promising, new development, but nevertheless there are still limitations to this voluntary approach. Most importantly, Donovan has highlighted how these arrangements do not solve the "unevenly distributed access" to data for researchers. She highlights how this creates "several problems for the advancement of social science, including fracturing development of a shared replication methodology, thus making evaluation of findings impossible. It also encourages secret transactions between researchers and companies, contradicting the scientific principles of openness, public good, and peer-review." [79]

III. DOES THE GDPR PREVENT PLATFORMS FROM PROVIDING ACCESS TO PLATFORM DATA TO INDEPENDENT RESEARCHERS?

UNTIL NOW, THIS PAPER has highlighted the valid reasons that independent researchers have to request access to platform data, including personal data. These requests have been acknowledged and supported by various officials and bodies of the EU equally interested in having access to platform data to understand platforms' role in facilitating and amplifying disinformation, as part of a new regulatory framework. Partly

as a result of this political pressure some platforms have provided voluntary access mechanisms to provide access to a limited subset of predominantly public or anonymized platform data, including via public APIs, or data-access grants and partnerships. These mechanisms are not seen as sufficient, both by researchers and European policymakers, for a variety of reasons.

One crucial reason a number of platforms have invoked for this limited appetite to expand the sharing of (personal) data is their obligations under the GDPR, which contains a basic premise: Personal data can only be processed by a "data controller" if the processing activity falls under one of six legal categories.[80] In this case, the data controllers are the platforms that organize, collate, and determine the means and purposes of processing.[81] By disseminating, transmitting, and making such data available to independent researchers, those data controllers will be engaging in an act of processing.[82]

A very strict interpretation of these obligations by the platforms argues that the GDPR would not allow the sharing of personal data from the platform to a researcher, while a more nuanced interpretation argues that the GDPR would not allow the sharing of personal data without taking precautions that in practice result in platforms only making available anonymous data or aggregated data. Whether those interpretations are true or not is subject to valid arguments on both sides. However, the fact that it requires an argument is in itself a barrier to hand over data. This uncertainty is undesirable for both researchers and the platforms.

This paper argues that the GDPR currently would allow platforms to share personal data based on consent or legitimate interest, and in the future, Article 31 of the DSA would allow the sharing of such data, and would in fact compel the sharing of such data for "vetted" academic researchers for the "purpose of conducting research that contributes to the identification and understanding of systemic risks," such as the dissemination of illegal content; inauthentic, automated, or otherwise intentional manipulation; and "negative effects" on the "exercise of fundamental rights."[83]

Consent as a legal basis to provide access to personal data

Personal data could be shared with independent researchers if the data subject has given his or her consent to the platform to share their personal data.[84]

Consent under the GDPR has specific requirements and a high threshold to be valid. The GDPR defines consent as "freely given, specific and informed, and granted by an unambiguous affirmative action."[85] That definition is augmented by specific "conditions for consent" in Article 7 of the GDPR. Note that one of the main bases for processing special category data is "explicit" consent.[86]

In addition to the threshold difficulties associated with consent, relying on consent could pose a number of other difficulties for platforms. For example, including a clause in the nonnegotiable terms of service of a platform that states that "by using this platform you agree that your data will be shared for research purposes" is unlikely to be "freely given" according to the GDPR, due to the power imbalance between the user and the platform, and hence a platform would likely not be able to rely on this form of consent as a legal basis to share personal data with an independent researcher.[87] Moreover, Article 7(2) of the GDPR means that consent cannot be bundled with other terms, and Article 7(3) of the GDPR means the data subject must be able to withdraw consent as easily as it was provided.

Furthermore, users of a specific platform would need to fully understand in advance the purpose of the research in order for their consent to be "specific and informed." This isn't always feasible—especially in the sort of big data research that characterizes the field of disinformation scholars.[88] Research usually happens *after* the data is collected, which makes it hard to provide "informed" consent a priori. The GDPR accommodates this reality of academic research via the research exemption in Article 5.1.b,[89] and users could in principle also consent to a platform sharing their data for more general research purposes[90] or on an ad hoc basis, where a platform asks a selection of their users whether they would want to contribute their data for a specific and well-delineated research project.

The practical implications of relying on consent, however, are not trivial. Researchers would need to be able to accommodate requests of users who withdraw their consent—which under the GDPR should be possible at any time during the research, and in a way that is as easy as giving consent.[91] This can have a significant impact on the research process and its conclusions. Requesting consent can also reduce the ultimate utility of a data set for researchers, since users that opt into a study may not be representative

of the relevant population the researcher is intending to study. In addition, users who consent in this way may change their behaviors precisely because they know those behaviors are being studied.

Legitimate interest as a legal basis to provide access to data

Personal data could also be shared for the purposes of the legitimate interests pursued by the platform or by the researcher.[92] In a 2017 case, the Court of Justice of the EU held that if a controller (in this case, the platform) wants to rely on legitimate interests as a legal basis, it has to satisfy a challenging three-part test.[93]

First, the platform has to demonstrate that the interest that is being pursued is legitimate. Carrying out scientific research is considered to be a legitimate interest,[94] but this interest must be "sufficiently clearly articulated to allow the balancing test to be carried out against the interests and fundamental rights of the data subject."[95] It could be argued that it would be difficult for platforms to balance these interests without having detailed knowledge in advance about the research purpose, aims, and analytical methods of a specific research request. However, requiring this information from a researcher in advance can jeopardize the independence of the researchers, and the legitimacy of any resulting study, since a platform would in theory be able to deny access if it does not like a particular topic of a research project.

Even less desirable, the platform would also need to show that a specific research methodology is "necessary" to achieve those aims.

Finally, the balancing test would require the platform to analyze to what extent the processing activity of the researcher might pose a risk to the fundamental rights of the data subject,[96] which would require inter alia knowledge of the procedural, contractual, and other safeguards that researchers are planning to put in place to safeguard the personal data and identities of the data subjects.

In summary, platforms could rely on legitimate interest as a legal basis to transfer data, but relying on this legal basis requires measures that could result in platforms acting as de facto gatekeepers that are able to decide on the validity of specific research proposals and methods, which would be

an undesirable situation. This concern also applies to reliance on consent as a legal basis for platforms to provide access to data, since the platforms would still need to assess the necessity and legitimacy of the purpose for the research design.

IV. HOW THE DIGITAL SERVICES ACT CREATES A NEW LEGAL OBLIGATION FOR PLATFORMS TO HAND OVER DATA TO RESEARCHERS

WHILE CONSENT AND LEGITIMATE interest are two legal mechanisms that platforms could use to provide access to data to researchers, both have problems. This section looks at another potential legal mechanism: A platform would be able to share data with independent researchers if this sharing is "necessary to comply with a legal obligation" to which the platform is subject in EU law or a national member state law.[97] There is currently no such EU-level legal obligation that would force platforms to share (personal) data for general research or auditing purposes, but the EU has envisaged such an obligation in Article 31 of the DSA. Article 31.2 states that "upon a reasoned request from the Digital Services Coordinator of establishment or the Commission, very large online platforms shall, within a reasonable period, as specified in the request, provide access to data to vetted researchers for the sole purpose of conducting research that contributes to the identification and understanding of systemic risks as set out in Article 26(1).[98] Let's unpack what this means.

The DSA will regulate the obligations of digital services that act as intermediaries in their role of connecting consumers with goods, services, and content. As such, it updates the e-Commerce Directive, which dated from 2000. According to the European Commission, these new rules are "an important step in defending European values in the online space," which will contribute to "setting a benchmark for a regulatory approach to online intermediaries also at the global level."[99] The proposals from the European Commission provide the starting point for both the European Parliament and the European member states to adopt legislation at the EU level. As co-legislators, they will first amend the proposals along with their preferences before agreeing on a

compromise text. This procedure is expected to last up until the summer of 2022 at the earliest.

The supervised risk assessment approach of the Digital Services Act

The most innovative part of the DSA proposal can be found in Chapter III, where the EU introduces a range of new due diligence obligations that are adapted to the type and nature of the intermediary service concerned. The proposal sets up a "supervised risk management approach,"[100] in which certain substantive obligations are imposed only to "very large platforms,"[101] which have an average of 45 million active users in the EU, and which due to their reach have acquired "a central, systemic role in facilitating the public debate and economic transactions."[102]

The European Commission argues that once a platform reaches this threshold, the systemic risks it poses can have a disproportionately negative impact on our societies given their reach and ability to facilitate public debate and disseminate information online. The European Commission argues that in the absence of effective regulation and enforcement, these platforms can design their services to "optimise their often advertising-driven business models," without "effectively identifying and mitigating the risks and the societal and economic harm they can cause."[103]

Hence, the European Commission imposes a duty on these platforms to conduct risk assessments on the systemic risks stemming from the functioning and use of their services, as well as potential misuses by the recipients of the services, and then take appropriate risk mitigating measures.[104] The proposal lists three broad categories of systemic risks, including (a) the dissemination of illegal content through their services, (b) any negative impact of the services on the exercise of fundamental rights, and (c) the intentional manipulation of their services that can have a foreseeable negative effect on a range of public policy goals. These include "the protection of public health, minors, civic discourse, or actual or foreseeable effects related to electoral processes and public security."[105] After having identified those risks, platforms should deploy "reasonable, proportionate and effective" means to mitigate those risks.[106] These include a broader range of actions than removing content and can include changes to algorithmic recommendation systems

or discontinuing advertising revenue for specific content.

Enforcement of the DSA will work on three different levels. The default rule is that ensuring adequate oversight and enforcement should, in principle, be attributed to the member states. However, the European member state in which the "main establishment" of the provider of the intermediary services is located, shall have jurisdiction over the due diligence obligations for platforms, which would typically be an Irish regulator. This regulator is called the "digital services coordinator of establishment."[107] Where systemic risks emerge across the EU posed by "very large online platforms," the proposed regulation provides for supervision and enforcement at the EU level, mainly via the European Commission.

These risk assessments and mitigating measures are subject to independent audits that can assess the effectiveness of these measures. Where the outcome of the audit is not positive, operational recommendations will be made regarding specific measures to achieve compliance with the company's obligations under the DSA.[108] Ultimately, if a platform doesn't sufficiently address these recommendations, the European Commission may further investigate this as a potential infringement of the DSA,[109] impose interim measures,[110] and ultimately impose fines of up to six percent of a company's total turnover in the preceding financial year.[111] Supplying "incorrect, incomplete or misleading information in response to a request" can lead to fines not exceeding one percent of the total turnover in the preceding financial year.[112]

Mandatory access for vetted researchers

The European Commission acknowledges that independent researchers can provide a crucial role in helping auditors, the "digital services coordinator of establishment," and the European Commission with their respective oversight roles to supervise whether platforms comply with their obligations under the DSA. Hence, the DSA provides a framework for "compelling access to data from very large online platforms to vetted researchers."[113]

Upon a "reasoned request" of the commission or the digital services coordinator of establishment, very large online platforms "shall provide access to data to vetted researchers … for the sole purposes of conducting research that contributes to the identification and understanding of systemic risks as set out in Article 26(1)."[114] Researchers need to fulfill four

main criteria in order to be "vetted" and be part of a pool of researchers that can undertake research on behalf of these actors. They need to: (i) be affiliated with academic institutions, (ii) be independent from commercial interests, (iii) have proven records of expertise in the fields related to the risks investigated or related research methodologies; and (iv) commit and be in a capacity to preserve the specific data security and confidentiality requirements corresponding to each request.[115] It is not clear how such a vetting process would work in practice, for instance through a professional licensing regime or via certification bodies.

The DSA includes a nonexhaustive list of three categories of data that should be provided by platforms,[116] through online databases or APIs,[117] including:

- the data necessary to assess the risks and possible harms brought about by the platform's systems,
- data on the accuracy, functioning, and testing of algorithmic systems for content moderation, recommender systems, or advertising systems, or
- data on processes and outputs of content moderation or of internal complaint-handling systems within the meaning of this regulation.

Article 31 is crucial for the EU's supervised risk management approach for three reasons: (1) researchers can act as a counterweight to the platforms' own risk assessment analysis, (2) researchers can act as pathfinders that dig up evidence of systemic risks, and (3) researchers can assess the effectiveness of platforms' proposed risk mitigating measures.

Researchers would get an unprecedented level of access to data, but the proposal takes into account legitimate interests of the platforms as well. A platform can ask to amend the request to provide data if it does not have access to the data or when giving access could in its view "lead to significant vulnerabilities for the security of its service or the protection of confidential information, in particular trade secrets."[118] However, given the crucial role for pre-vetted researchers to identify systemic risks that need to be mitigated, and their role as independent information providers to auditors,[119] this derogation is overly broad. Recital 60 of the DSA states explicitly that very

large online platforms should give the auditor "access to all relevant data necessary to perform the audit properly" and the auditors should also "be able to make use of other sources of objective information, including studies by vetted researchers." It further says that auditors should "guarantee the confidentiality, security and integrity of information, such as trade secrets." Pre-vetted researchers must live up to the same standards, and their vetting process should be conditional upon their ability to live up to those standards, but security reasons and trade secrets should not be a ground for a platform to refuse access to data a priori.

The proposal also states that all requirements for access to data under this framework should be "proportionate and appropriately protect the rights and legitimate interests, ... including the recipients of the service."[120] As Donovan points out, "researchers can unintentionally expose research subjects to a range of harms, including identity theft, financial fraud, harassment, abuse, or reidentification."[121] Potential GDPR violations can also be added to this list. This is where potential GDPR concerns about sharing of data enter the DSA framework. The European Commission stated in its European Democracy Action Plan from December 2020 that "the GDPR does not a priori and across the board prohibit the sharing of personal data by platforms with researchers,"[122] and opened the door in the DSA to solve these challenges at a later stage. It further announced that it would adopt a separate delegated act to "lay down the technical conditions under which platforms should share data and the purposes for which the data may be used. This delegated act will elaborate the "specific conditions under which such sharing of data with vetted researchers can take place" in compliance with the GDPR.[123]

One point of criticism to this proposal is that it limits access to data to *academic* researchers. From a GDPR-perspective, it does make sense to limit access to data to *scientific* researchers, but this does not necessarily need to be translated as academic researchers. This is one issue that can be solved potentially through a code of conduct on GDPR-compliant access to platform data.

Establish a code of conduct on GDPR-compliant access to platform data to facilitate sharing of data

The DSA would provide a much-welcomed legal basis for platforms to hand over access to platform data to academic researchers. However, this framework for compelling platforms to make data available does not solve the question of how researchers could get access outside this specific framework, without the intervention of the European Commission or the digital services coordinator of residence. Also, as the Commission already acknowledged, establishing this legal basis is only a first step to enable GDPR-compliant access to data.

Once that legal basis is known, the platform — in its role as controller — also needs to take into account a number of key data protection principles before it can share data. These principles include that the platform has to ensure that the data is shared in accordance with the purposes for which they were collected ("purpose limitation"),[124] ensure that technical and organizational measures are put in place in order to respect the principles of data minimization,[125] accuracy,[126] storage limitation,[127] and preservation of the integrity and confidentiality of the data.[128] The platforms also would need to adhere to a number of transparency obligations.[129]

The drafters of the GDPR implicitly conceded that adhering to all these principles might be difficult in a research context, and hence introduced a number of derogations to some of these principles. Importantly, the sharing of special category data may be allowed if this is seen as necessary for archiving purposes in the public interest, scientific, or historical research purposes, if the processing complies with the requirements imposed by Article 89 (1) of the GDPR.[130] Additionally, there are three other crucial derogations: (1) the GDPR presumes that sharing data for research purposes is considered to be compatible with the initial purposes of data collection,[131] (2) personal data used for research may be exempt from the exercise of individual rights, such as the right to request access to your data,[132] and (3) platforms are exempt from providing data subjects with specific pieces of information.[133]

While the GDPR sets out the general principles and rights that need to be respected by data processors and data controllers, it does not spell out in detail how these rights and principles should be applied in very specific

contexts. Neither does the DSA. To solve this situation, the GDPR has encouraged the creation of codes of conduct under Article 40 of the GDPR, where associations and other bodies representing categories of controllers or processors can "contribute to the proper application" of the GDPR, "taking account of the specific features of the various processing sectors."[134] Such a code should aim to codify how the GDPR shall apply in a "specific, practical and precise manner."[135]

Associations of academics in a specific field, trade associations that represent platforms, consortia of associations, or any other body that represents a relevant sector in this field may (jointly) prepare such codes; they become the "code owners."[136] Together these actors can develop and agree upon technical standards, logistical protocols and joint interpretations of the GDPR that demonstrate how both the platform and the researcher comply with specific parts or principles of the GDPR. In specifying the application of the GDPR provisions to a processing activity or a sector, the code should provide sufficient added value by using, for example, sector-specific terminology and offering use cases and best practices.

The European data protection supervisor (EDPS) encouraged the adoption of a code of conduct for facilitating access by researchers. The EDPS specifically added that such a code can include "the provision by private companies, particularly tech platforms, of data to independent researchers for specific projects, such as examining online manipulation and the dissemination of misinformation."[137] Similarly, the European Commission has been keen to support "an effective data disclosure for research on disinformation … by developing a framework in line with applicable regulatory requirements and based on the involvement of all relevant stakeholders (and independent from political influence)."[138]

Importantly, the drafters of this code need to submit a draft to the relevant national data protection authority,[139] which needs to approve it, provided that it complies with the GDPR. For transnational codes, which cover data processing activities in the territory of multiple EU member states, the competent supervisory authority must subsequently refer the code to the European Data Protection Board (EDPB) for a decision,[140] and from there, it proceeds to the European Commission,[141] which can decide that the approved code of conduct has general validity within the EU. The code

of conduct then ends up as an implementing act in EU law.[142]

Monitoring adherence to the code: Limited mandates vs. broader mandates

One of the most challenging aspects of establishing a code in this context is that Article 41 of the GDPR requires the establishment of a monitoring body, which scrutinizes the compliance of the platforms and research associations with the code and which carries out reviews of the code's operations. The EDPB specifies that this body needs to have at its disposal effective oversight mechanisms, including "random and unannounced audits, annual inspections,"[143] "reporting requirements, clear and transparent complaint handling and dispute resolution procedures, concrete sanctions and remedies in cases of violations of the code, as well as policies for reporting breaches of its provisions."[144] Importantly, the monitoring body can sanction those members who break the rules of the code.[145]

Such a body can have a limited function, and only act as a monitoring body, but in theory, it could also be part of a larger organization with a broader mandate. Such "trusted third parties" could provide a secure environment in which companies can make personal data available to trusted—or vetted—third parties for further processing for research purposes. These don't necessarily need to be limited to academic researchers. The Belgian implementation of the GDPR allows recourse to such a trusted third party if it is independent from both the initial and the subsequent controllers.[146] Similar mechanisms also exist in the U.K.[147] and in France, where the Centre for Secure Access to Data focuses on organizing and implementing secure access services for confidential data for nonprofit research, study, evaluation, or innovation.[148]

Jef Ausloos and others correctly identify the strengths of such an independent institution, which can act as a bridge between those holding the data and those wishing to get access to it. Such a trusted third party can act as a "neutral arbiter" in deciding on requests for confidentiality from the disclosing party and in periodically auditing disclosing parties to verify the accuracy of disclosures."[149] It can "maintain relevant access infrastructure" and "verify and pre-process corporate data in order to ensure it is suitable for disclosure."[150] Ultimately, it could even play a role in "ensuring otherwise

unavailable or uninterpretable data to be made accessible" because "the fact that data is not readily available and/or produced by the respective platforms, should not be a reason to discard including that data into the access regime." In the platform context, Ausloos and others suggest that a "more centralized EU-level institution"[151] could be advisable given "the political-economic power and multinational dimensions of platform operators."[152]

Topics that could be clarified in a code of conduct on access to platform data

Currently, platforms are reluctant to hand over personal data, let alone special category data, to independent researchers because they perceive a number of uncertainties related to the research exemptions of the GDPR in particular. In the following paragraphs, this essay will attempt to highlight five topics which could be clarified in a code of conduct.

PROVIDE CLARITY ON WHICH ACTORS CAN QUALIFY AS RESEARCHERS

One obstacle that some platforms have brought up in discussions is that they should not be placed in a position to define whether a specific activity qualifies as "research" under the GDPR. The GDPR does not explicitly define research but simply states that it should be interpreted "in a broad manner," including, for example, "technological development and demonstration, fundamental research, applied research and privately funded research."[153] The EU has always used a broad interpretation of the notion of research, illustrated by Article 179 of the Treaty on the Functioning of the European Union that specifies that "the Union shall have the objective of strengthening its scientific and technological bases by achieving a European research area in which researchers, scientific knowledge and technology circulate freely."

Both the EDPS and the EDPB have clarified that the application of the special data protection regime for scientific research applies to research where "relevant sectoral standards of methodology and ethics apply,"[154] or where a research project is "set up in accordance with relevant sector-related methodological and ethical standards, in conformity with good practice."[155] The code could lay out what those specific methodological and ethical standards are, and clarify whether affiliation with an academic institution is a

prerequisite to meet those standards. This is not only relevant from a GDPR perspective but from a broader equity perspective. The current voluntary mechanisms to share data tend to privilege established actors working in this space, which creates asymmetries among institutions and researchers that are likely to only increase over time.

PROVIDE CLARITY ON SAFEGUARDS TO RECEIVE PLATFORM DATA ...

Under Article 5(1)(b) of the GDPR, researchers can receive data that were initially collected by the platform for different purposes, in order to subsequently use these for research. However, to benefit from this relaxed rule the receiving researcher needs to implement "appropriate safeguards" to protect the rights and freedoms of the data subject.[156] According to the EDPS, this doesn't provide "a general authorization" to further process data in all cases for historical, statistical or scientific purposes. Instead, "each case must be considered on its own merits and circumstances."[157] Some platforms have argued that they don't want to be the arbiter that decides which types of technical, organizational, procedural, or contractual safeguards should be implemented in a research project before data can be shared. This is a legitimate concern. For example, the EDPB has specified that in scientific research, data minimization can be achieved "through the requirement of specifying the research questions and assessing the type and amount of data necessary to properly answer these research questions."[158] Platforms shouldn't be making that assessment as it could have an impact on the independence of the research.

... AND SEND PLATFORM DATA

Anonymization of personal data is encouraged by the GDPR as a way to mitigate risks for the data subject before data is shared.[159] Some platforms have argued that the guidance from the Article 29 Data Protection Working Party on anonymization techniques contains contradictions and is confusing. On the one hand, the document sets a very high bar to achieve anonymity; it states that data are seen as anonymous data if the process to strip the data of identifiable elements is "irreversible."[160] At the same time, however, it also allows for a lower bar, suggesting that an anonymization

process "is sufficiently robust" if "identification has become reasonably impossible."[161] A code of practice could clarify how these two standards could be seen as mere examples of how anonymization can be achieved, or can highlight which anonymization techniques are best suited for specific research purposes.

CLARIFY TRANSPARENCY REQUIREMENTS FOR DATA SUBJECTS

The GDPR requires platforms to provide their users with relevant information about the purposes of the processing of their data and the recipients of any personal data collected.[162] Despite an explicit GDPR exception to that principle when further processing is for research purposes, some platforms are uncertain that these provisions apply to them. A code of conduct could clarify the extent, timing, and scope of the information about the research protocol that needs to be given to both the data subjects and the broader public. Moreover, the code could cover the extent to which data subjects can exercise their rights in the context of the potential limitations of those rights.

CLARIFY THE APPLICATION OF DATA SUBJECT RIGHTS

One of the most challenging aspects related to the GDPR, in general, is that it has left quite some discretion to the member states of the EU to implement the law. Different member states have different rules regarding how special-category data and personal data are allowed to be processed for research purposes.[163] The research exceptions that exclude the application of some of the data subject rights, such as rights to access or rectification, can also vary from member state to member state.[164] This means in practice that personal data that was made available by a platform for research purposes could be subject to different regimes for individuals to exercise their rights, depending on which member state law applies. The code could provide guidance or best practices on this topic as well.

CONCLUSION

The EU is a trailblazer in proposing a legally binding obligation for platforms to provide access to platform data to independent researchers as part of its ongoing legislative efforts to establish additional responsibilities for online platforms. Such an obligation is necessary to understand the role and impact of platforms in our societies, which in turn can enable evidence-based policies to address perceived systemic risks to our societies. The need for greater access to data is illustrated in this essay by focusing on one specific key policy discussion: the role of social media platforms in enabling, facilitating, and amplifying disinformation.

However, this framework for compelling platforms to make data available does not solve the question of how researchers could get access outside the specific framework of the DSA, and more guidance is needed to allow the sharing of data in a privacy-proof manner. The adoption of a code of conduct on access to platform data in the EU could provide more legal clarity in this area, as it would facilitate the interpretation of the GDPR in this specific context for both researchers and platforms. It has the additional benefit that its work can have a potential impact for researchers outside the EU, as it can provide a source of inspiration to the broader field of disinformation researchers. Finally, the European Commission's approach—building data-access requirements into broader platform transparency measures—holds significant potential in the U.S. as well. This strategy is flexible enough to allow data access provisions to be incorporated into almost any federal legislative package related to digital platforms—from data privacy to election integrity, cybersecurity to online advertising legislation. While the EU code of conduct is based on an article within the GDPR, which is without parallel in the U.S., the U.S. has a long-standing tradition of incorporating independent codes of conduct into federal and state regulatory frameworks.

NOTES

1 Gary King & Nate Persily, *Building Infrastructure for Studying Social Media's Role in Elections and Democracy*, Social Science One, 27 August 2019, https://socialscience.one/blog/building-infrastructure-studying-social-media's-role-elections-and-democracy%C2%Ao [https://perma.cc/X9E5-HL79].

2 Mathias Vermeulen, *The keys to the kingdom. Overcoming GDPR-concerns to unlock access to platform data for independent researchers*, OSF Preprints, 27 November 2020, https://osf.io/vnswz/ [https://perma.cc/DPP7-292L].

3 European Commission, *Proposal for a Digital Services Act (DSA)*, 15 December 2020, https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN [https://perma.cc/9L7D-6WTZ].

4 Sinan Aral & Dean Eckles, *Protecting elections from social media manipulation*, 365 Science 858, 2019, https://doi.org/10.1126/science.aaw8243 [https://perma.cc/8DJ5-9RZS].

5 Paddy Leerssen, *The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems*, 11 European Journal of Law and Technology 2, 24 February 2020, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3544009 [https://perma.cc/HQ24-PYAE].

6 General Data Protection Regulation, GDPR art. 89.

7 A/B testing consists of a randomized experiment with two variants, A and B, in which two versions of a single variable are tested, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

8 *See e.g.*, Mozilla, *Facebook and Google: This is What an Effective Ad Archive API Looks Like*, Mozilla Blog, 28 March 2019, https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like/ [https://perma.cc/7FRA-YBLB].

9 *See e.g.*, Brandi Geurkink, *Our Recommendation to YouTube*, Mozilla, 14 October 2019, https://foundation.mozilla.org/en/blog/our-recommendation-youtube/ [https://perma.cc/F6TX-R24J].

10 European Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services § 24: "The notion of main parameter should be understood to refer to any general criteria, processes, specific signals incorporated into algorithms or other adjustment or demotion mechanisms used in connection with the ranking."

11 *Id.*

12 GDPR, *supra* note 6, at art. 4.1.

13 *Id.*

14 *Id.* at art. 9.1.

15 *Id.*

16 Joan Donovan, *Redesigning consent: Big data, bigger risks*, Harvard Kennedy School Misinformation Review, 14 January 2020, https://misinforeview.hks.harvard.edu/article/big-data-bigger-risks/ [https://perma.cc/U388-UEPK].

17 *See also* Thomas Ridt, D*isinformation: A Primer in Russian Active Measures and Influence Campaigns*, Hearings before the Select Committee on Intelligence United States Senate, 30 March 2017, https://www.intelligence.senate.gov/sites/default/files/documents/os-trid-033017.pdf [https://perma.cc/2BBY-W5NP].

18 Camille Francois, *Actors, Behaviors, Content: A Disinformation ABC*, Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, 20 September 2019, https://science.house.gov/imo/media/doc/Francois%20Addendum%20to%20Testimony%20-%20ABC_Framework_2019_Sept_2019.pdf [https://perma.cc/8397-49YJ].

19 *Id.*

20 Camille François, *Briefing for the United States House of Representatives Committee on Science, Space, and Technology*, Graphika, p. 4, 26 September 2019, https://science.house.gov/imo/media/doc/Francois Testimony.pdf [https://perma.cc/7WH6-K8Q2].

21 James Pamment, *EU Code of Practice on Disinformation: Briefing Note for the New European*

Commission, Carnegie Endowment for International Peace, 3 March 2020, https://carnegieendowment.org/2020/03/03/eu-code-of-practice-on-disinformation-briefing-note-for-new-european-commission-pub-81187 [https://perma.cc/6NCT-L7UJ].

22 Irene V. Pasquetto et al., *Tackling misinformation: What researchers could do with social media data*, Harvard Kennedy School Misinformation Review, 9 December 2020, https://misinforeview.hks.harvard.edu/article/tackling-misinformation-what-researchers-could-do-with-social-media-data/ [https://perma.cc/93KC-7E5L].

23 *Id*. at Soroush Vosoughi, *The need for impression data in misinformation research*.

24 *Id*. at Miriam Metzger & Andrew J. Flanagi, *Understanding how communication about misinformation can help to combat it*.

25 *Id*. at Brian E. Weeks, *Emotion, social media and misinformation*.

26 *Id*. at Jasmine McNealy & Seungahn Nah, *Misinformed citizens across social media platforms: Unraveling the effects of misinformation on social capital and civic participation*.

27 Nick Clegg & Chaya Nayak, *New Facebook and Instagram Research Initiative to Look at US 2020 Presidential Election*, Facebook, 31 August 2020, https://about.fb.com/news/2020/08/research-impact-of-facebook-and-instagram-on-us-election/ [https://perma.cc/QL3L-2SZF].

28 Zeynep Tufekci, *It's the (Democracy-Poisoning) Golden Age of Free Speech*, Wired, 16 January 2018, https://www.wired.com/story/free-speech-issue-tech-turmoil-new-censorship/ [https://perma.cc/C8YA-2WR6]. *See also* Facebook co-founder Chris Hughes' statement: "Facebook's business model is built on capturing as much of our attention as possible to encourage people to create and share more information about who they are and who they want to be," from *It's Time to Break Up Facebook*, New York Times, 9 May 2019, https://www.nytimes.com/2019/05/09/opinion/sunday/chris-hughes-facebook-zuckerberg.html [https://perma.cc/FE6P-HCMY].

29 Christina Nemr & William Gangware, *Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age*, Park Advisors, March 2019, https://www.state.gov/wp-content/uploads/2019/05/Weapons-of-Mass-Distraction-Foreign-State-Sponsored-Disinformation-in-the-Digital-Age.pdf [https://perma.cc/B5NQ-7K7Y], at p. 36. *See also Final report of the High Level Expert Group on Fake News and Online Disinformation*, European Union, 12 March 2018, https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation [https://perma.cc/3CJY-HCU5], at p. 12, stating that it "is clear that many of the tools integral to the contemporary digital ecosystem that are used for legitimate purposes— e.g., behavioural data collection, analytics, advertising exchanges, tools for cluster detection and tracking social media sentiment, and various forms of AI/machine learning—have also been harnessed by some purveyors of disinformation."

30 Chloe Colliver & Jennie King, *The First 100 Days: Coronavirus and Crisis Management on Social Media Platforms*, Institute for Strategic Dialogue, 11 June 2020, https://www.isdglobal.org/wp-content/uploads/2020/06/First-100-Days.pdf [https://perma.cc/5Z8B-NUAY].

31 *Id*.

32 Charlie Warzel, *QAnon Was a Theory on a Message Board. Now It's Headed to Congress*, New York Times, 15 August 2020, https://www.nytimes.com/2020/08/15/opinion/qanon-marjorie-greene-congress.html [https://perma.cc/JK8A-GFZL].

33 Jeff Horwitz & Deepa Seetharman, *Facebook Executives Shut Down Efforts to Make the Site Less Divisive*, Wall Street Journal, 26 May 2020, https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499?mod=hp_lead_pos5 [https://perma.cc/FX53-JZQS].

34 Aaron Sankin, *Want to Find a Misinformed Public? Facebook's Already Done It*, The Markup, 23 April 2020 https://themarkup.org/coronavirus/2020/04/23/want-to-find-a-misinformed-public-facebooks-already-done-it [https://perma.cc/9JX5-949H].

35 Donna Lu, *Facebook's fact-checking process is too opaque to know if it's working*, New Scientist, 30 July 2019, https://www.newscientist.com/arti-

cle/2211634-facebooks-fact-checking-process-is-too-opaque-to-know-if-its-working/ [https://perma.cc/GZ6F-MTFV].

36 Pasquetto et al., *supra* note 22, at Gordon Pennycook & David. G. Rand, *Evaluating interventions to fight misinformation*.

37 European Commission, *supra* note 29. As an example, the report highlights how some platform companies, like Facebook and Google, "have experimented with flagging links to disinformation on their products and services or highlighting fact-checking work. To evaluate the effectiveness of such responses, platform companies should set out a clear timeline for sharing information on how many links are flagged, what percentage of overall content surfaced this represents, and other data that allows independent third-parties including fact-checkers and researchers to evaluate impact. To ensure these responses do not have adverse consequences, it is also important that they work on appeals processes and make clear how these appeals work."

38 Pasquetto et al., *supra* note 22, at Michelle A. Amazeen & Chris Vargo, *The influence of fact checks and advertising*.

39 Colliver & King, *supra* note 30, at p. 5. *See also* Nathalie Maréchal et al., *Getting to the Source of Infodemics: It's the Business Model, New America*, May 2020, https://www.newamerica.org/oti/reports/getting-to-the-source-of-infodemics-its-the-business-model/ [https://perma.cc/ADK8-S3N6].

40 Jef Ausloos et al., *Operationalizing Research Access in Platform Governance: What to Learn from Other Industries?* Algorithm Watch, p. 17, 25 June 2020, https://algorithmwatch.org/wp-content/uploads/2020/06/GoverningPlatforms_IViR_study_June2020-AlgorithmWatch-2020-06-24.pdf [https://perma.cc/8VG4-AXR3].

41 Zeynep Tufekci, *Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls*, Eighth International AAAI Conference on Weblogs and Social Media, 28 March 2014, https://arxiv.org/abs/1403.7400 [https://perma.cc/8Q8K-U6Y7].

42 Axel Bruns, *After the 'APIcalypse': social media platforms and their fight against critical scholarly research*, Information, Communication & Society, 22:11, p. 1544-1566, 11 July 2019, https://www.tandfonline.com/doi/abs/10.1080/1369118X.2019.1637447 [https://perma.cc/S4R2-MYP9].

43 Alexander Sangerlaub, *The blind spot of digital publics*, Stiftung Neue Verantwortung, 21 March 2019, https://www.stiftung-nv.de/de/publikation/der-blinde-fleck-digitaler-offentlichkeiten [https://perma.cc/NRE8-BGF9].

44 Matt Garmur et al., *CrowdTangle Platform and API*, Harvard Dataverse, 23 January 2019, V3, p. 8, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/SCCQYD [https://perma.cc/VP9G-2XBE]. For instance, the "benchmark comments" are the expected number of comments a post should have after a given amount of time." are the expected number of comments a post should have after a given amount of time.

45 Bruns, *supra* note 42.

46 European Commission, *COM/2018/236 final, Tackling online disinformation: a European Approach*, 26 April 2018, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0236 [https://perma.cc/88AK-4AR6].

47 *Id*.

48 European Commission, *Code of Practice on Disinformation*, 28 May 2021, https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation [https://perma.cc/FWJ9-6CEK].

49 *Id*. (emphasis added).

50 *Id*.

51 *Id*.

52 All elements of the Ad Library are defined in this document, Erika Franklin Fowler et al., *Ad Library API Code Book*, 25 April 2019, https://social-science.one/files/ad_library_api_codebook.pdf [https://perma.cc/ZKQ9-ZYA4].

53 Mozilla, *supra* note 8.

54 Natasha Lomas, *Facebook accused of blocking wider efforts to study its ad platform*, TechCrunch, 29 April 2019, https://techcrunch.com/2019/04/29/facebook-accused-of-blocking-wider-efforts-to-study-its-ad-platform/ [https://perma.cc/DKJ8-5JHB].

55 Matthew Rosenberg, *Ad Tool Facebook Built to Fight Disinformation Doesn't Work as Advertised*,

New York Times, 25 July 2019, https://www.nytimes.com/2019/07/25/technology/facebook-ad-library.html [https://perma.cc/6PFL-PLUP].

56 *Id.*

57 *See* French Ambassador for Digital Affairs, *Facebook Ads Library Assessment*, https://disinfo.quaidorsay.fr/en/facebook-ads-library-assessment [https://perma.cc/6PFL-PLUP]. For an extensive overview of flaws relating to the Ad library API see Mozilla, *Data Collection Log — EU Ad Transparency Report*, https://adtransparency.mozilla.org/eu/log/ [https://perma.cc/642G-NDCS]. For an overview of the effectiveness of the ad library in the 2019 Italian, Czech, and Dutch elections for the European Parliament see European Partnership for Democracy, *Virtual Insanity? The need to guarantee transparency in digital political advertising*, March 2020, https://epd.eu/wp-content/uploads/2020/04/Virtual-Insanity-synthesis-of-findings-on-digital-political-advertising-EPD-03-2020.pdf [https://perma.cc/4XZG-MDXH].

58 Jeff Horwitz, *Facebook Seeks Shutdown of NYU Research Project Into Political Ad Targeting*, Wall Street Journal, 23 October 2020, https://www.wsj.com/articles/facebook-seeks-shutdown-of-nyu-research-project-into-political-ad-targeting-11603488533 [https://perma.cc/9D7B-R6DK].

59 Bloomberg Quicktake, Twitter, 30 October 2019, https://twitter.com/QuickTake/status/1189731760571072513?s=20 [https://perma.cc/FR5A-SAZ4].

60 European Commission, *First monthly intermediate results of the EU Code of Practice against disinformation*, 8 March 2021, https://digital-strategy.ec.europa.eu/en/news/first-monthly-intermediate-results-eu-code-practice-against-disinformation [https://perma.cc/XA8J-NVUA].

61 Julian King, *Remarks at the Press Conference for the 18th Security Union Progress Report*, 20 March 2019, https://web.archive.org/web/20191129195403/https://ec.europa.eu/commission/commissioners/2014-2019/king/announcements/commissioner-kings-remarks-press-conference-18th-security-union-progress-report-and-progress-made_en.

62 This assessment was based on two other audit reports that came to the same conclusions. *See* ERGA, *ERGA Report on Disinformation: Assessment of the implementation of the Code of Practice*, 2019, http://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf [https://perma.cc/TA7V-KUB2].

63 European Commission Staff Working Document, *Assessment of the Code of Practice on Disinformation – Achievements and areas for further improvement*, p. 12, https://digital-strategy.ec.europa.eu/en/library/assessment-code-practice-disinformation-achievements-and-areas-further-improvement [https://perma.cc/JY38-98LN]. *See also* European Commission, *Study for the assessment of the implementation of the Code of Practice on Disinformation*, 8 May 2020, https://ec.europa.eu/digital-single-market/en/news/study-assessment-implementation-code-practice-disinformation [https://perma.cc/S729-PJLC].

64 Social Science One, https://socialscience.one/ [https://perma.cc/UA3X-HJAA]. Gary King & Nathaniel Persily, *A New Model for Industry—Academic Partnerships*, 54 Political Science and Politics 3, p. 1, 2019, https://gking.harvard.edu/files/gking/files/partnerships.pdf [https://perma.cc/3USC-W2BP].

65 Social Science One, *Public Launch*, 11 July 2018, https://socialscience.one/blog/social-science-one-public-launch [https://perma.cc/X3PX-4TG7].

66 Letter to the Social Science Research Council from Funders Supporting Independent Scholarly Access to Facebook Data, 27 August 2019, https://ssrc-static.s3.amazonaws.com/sdi/resources/SM-DRG_funder_letter_august_2019.pdf [https://perma.cc/NR7V-DCDE].

67 Social Science One, *Public statement from the Co-Chairs and European Advisory Committee of Social Science One*, 11 December 2019, https://socialscience.one/blog/public-statement-european-advisory-committee-social-science-one [https://perma.cc/L2R7-EBAH].

68 Elizabeth Gibney, *Privacy hurdles thwart Facebook democracy research*, Nature, 3 October 2019, https://www.nature.com/articles/d41586-019-02966-x [https://perma.cc/MUC2-FB9P].

69 European Commission, *Facebook Reports on Implementation of the Code of Practice on Disinformation*, April 2019, https://ec.europa.eu/newsroom/

dae/document.cfm?doc_id=59225 [https://perma.cc/QPH5-EG2G].

70 FTC, *Facebook Settles FTC Charges That It Deceived Consumers By Failing To Keep Privacy Promises*, 29 November 2011, https://www.ftc.gov/news-events/press-releases/2011/11/facebook-settles-ftc-charges-it-deceived-consumers-failing-keep [https://perma.cc/JLW6-T3YU]. In November 2011, Facebook agreed to settle Federal Trade Commission charges that it deceived consumers by telling them they could keep their information on Facebook private, and then repeatedly allowing it to be shared and made public. The proposed settlement requires Facebook to take several steps to make sure it lives up to its promises in the future, including obtaining consumers' express consent before their information is shared beyond the privacy settings they have established.

71 Gary King & Nathaniel Persily, *Unprecedented Facebook URLs Dataset now Available for Academic Research through Social Science One*, Social Science One, 13 February 2020, https://socialscience.one/blog/unprecedented-facebook-urls-dataset-now-available-research-through-social-science-one [https://perma.cc/5HSD-5TWV].

72 Social Science One, *supra* note 67.

73 King & Persily, *supra* note 71.

74 *Id*.

75 *Id*.

76 Clegg & Nayak, *supra* note 27.

77 Talia Stroud et al., *A Proposal for Understanding Social Media's Impact on Elections: Rigorous, Peer-Reviewed Scientific Research*, 2020 Election Research Project, 31 August 2020, https://medium.com/@2020_election_research_project/a-proposal-for-understanding-social-medias-impact-on-elections-4ca5b7aae10 [https://perma.cc/TW9B-P9MR].

78 *Id*.

79 Donovan, *supra* note 16.

80 GDPR, *supra* note 6, at art. 5.1.a. The GDPR creates a higher threshold of protection for the processing of "special category" personal data, which includes data revealing political opinions or religious or philosophical beliefs. GDPR art. 9.1.

81 *Id*. at art. 4.7.

82 *Id*. at art. 4.2.

83 DSA, *supra* note 3, at art. 31.

84 *Id*. at art. 6.1.a.

85 *Id*. at art. 4.11.

86 *Id*. at art. 9.2.a.

87 European Data Protection Board, *Guidelines 05/2020 on consent under regulation 2016/679*, Version 1.1, p.13, 4 May 2020, https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf [https://perma.cc/TNH3-TFUE].

88 The Association of Internet Researchers states in its ethical guidelines that informed consent is "manifestly impracticable in the case of big data projects," which raises "serious ethical dilemmas," Association of Internet Researchers, Internet Guidelines 3.0, p.10, 6 October 2019, https://aoir.org/reports/ethics3.pdf [https://perma.cc/F98Q-6AEX].

89 GDPR, *supra* note 6, at art. 5.1.b states that further processing of data for scientific purposes shall not be considered to be incompatible with the purpose limitation principle.

90 *Id*. at recital 33. *See also* European Data Protection Board, *supra* note 87, at p. 30.

91 *Id*. at art. 7.3.

92 *Id*. note 6 at art. 6.1.f.

93 CJEU Case C-13-16, Valsts policijas Rīgas reģiona pārvaldes Kārtības policijas pārvalde v Rīgas pašvaldības SIA 'Rīgas satiksme', 4 May 2017 at para. 28, https://curia.europa.eu/juris/document/document.jsf?text=&docid=190322&doclang=EN.

94 The Article 29 Working Party was an advisory body made up of a representative from the data protection authority of each EU Member State, the European data protection supervisor, and the European Commission, which interpreted the Data Protection Directive, the predecessor of the GDPR. Its successor is the European Data Protection Board (EDPB). The Article 29 Working Party stipulated that a "well-considered use" of legitimate interest as a legal basis to provide access to data is especially relevant for "historical or other kinds of scientific research, particularly where access is required

to certain databases." Article 29 Data Protection Working Party, *Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC*, 9 April 2014, https://www.dataprotection.ro/servlet/ViewDocument?id=1086 [https://perma.cc/VYG2-EA4D].

95 *Id*. at p. 24.

96 GDPR, *supra* note 6, at recital 47.

97 *Id*. at art. 6.1.c.

98 DSA, *supra* note 3, at art. 31.2.

99 European Commission, *Digital Services Act—Questions and Answers*, 15 December 2020, https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348 [https://perma.cc/6JWH-3C4R].

100 DSA, *supra* note 3, at p. 11.

101 *Id*. at art. 25.1. These are platforms that provide their services to a number of average monthly active recipients of the service in the EU equal to or higher than 45 million (roughly 10 percent of the EU's population).

102 *Id*. at §53.

103 *Id*. at §56.

104 *Id*. at art. 26 and §57.

105 *Id*. at art. 26.

106 *Id*. at art. 27.1, §58.

107 *Id*. at art. 2.l.

108 *Id*. at art. 28.3.f.

109 *Id*. at art. 51.1.c.

110 *Id*. at art. 55.

111 *Id*. at art. 59.1.

112 *Id*. at art. 59.2.

113 *Id*. at §64.

114 *Id*. at art. 31.2.

115 *Id*. at art. 31.4.

116 *Id*. at §64.

117 *Id*. at art. 31.3.

118 *Id*. at art. 31.6.b.

119 *Id*. at §60.

120 *Id*. at §64.

121 Donovan, *supra* note 16.

122 European Commission, *On the European democracy action plan*, 3 December 2020, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423 [https://perma.cc/Q2U2-48Z3].

123 DSA, *supra* note 3, at art. 31.5.

124 GDPR, *supra* note 6, at art. 5.1.b.

125 *Id*. at art. 5.1.c.

126 *Id*. at art. 5.1.d.

127 *Id*. at art. 5.1.e.

128 *Id*. at art. 5.1.f.

129 *Id*. at art. 13.

130 *Id*. at art. 89.1. "Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner."

131 *Id*. at art. 5.1.b.

132 *Id*. at art. 17.3.d and art. 89.2.

133 *Id*. at art. 14.5.b.

134 *Id*. at art. 40.1.

135 European Data Protection Board, *Guidelines 1/2019 on Codes of Conduct and Monitoring Bodies under Regulation 2016/679*, V 2.0, p. 15, 4 June 2019, https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_201901_v2.0_codesofconduct_en.pdf [https://perma.cc/GS3M-8HJ2].

136 *Id*. at p. 7. The EDPB specifies that the code owners must demonstrate that "they are an effective representative body and that they are capable of understanding the needs of their members and clearly defining the processing activity or sector to which the code is intended to apply." *Id*. at. p. 11-12.

137 European data protection supervisor, *A preliminary opinion on data protection and scientific research*, p. 26, 6 January 2020, https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf [https://perma.cc/4TBM-X3S4].

138 European Data Action Plan, *supra* note 122.

139 GDPR *supra* note 6, at art. 40.5. The EDPB lists a number of factors that could take into account to decide which data processing agreement is relevant: (1) The location of the largest density of the processing activity or sector; (2) the location of the largest density of data subjects affected by the processing activity or sector; (3) the location of the code owner's headquarters; (4) the location of the proposed monitoring body's headquarters; or (5) the initiatives developed by a supervisory authority in a specific field. EDPB, *supra* note 135, at p. 28.

140 GDPR, *supra* note 6, at art. 40.7.

141 *Id*. at art. 40.8.

142 *Id*. at art. 40.9.

143 EDPB, *supra* note 135, at p. 23.

144 *Id*. 1332 at p. 16.

145 GDPR, *supra* note 6, at art. 41.2.b and c.

146 Art. 202 of the Belgian Law implementing the GDPR "*Wet van 30 juli 2018 betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens.*"

147 *See* UK Statistics Authority, National Statistician's Data Ethics Advisory Committee, https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ [https://perma.cc/UAT9-HW4J].

148 For more information, see CASD (The Secure Access Data Center), https://www.casd.eu/en/ [https://perma.cc/GVB8-Q4BT].

149 Ausloos et al., *supra* note 40, at p. 84.

150 *Id*.

151 *Id*. at p. 87-88.

152 *Id*. at p. 84.

153 GDPR, *supra* note 6, at recital 159.

154 EDPS, *supra* note 135, at p. 10-12.

155 EDPB, *supra* note 87, at p. 143.

156 GDPR, *supra* note 6, at art. 89.1.

157 EDPS, *supra* note 135, at p. 22.

158 EDPB, *Guidelines 03/2020 on the processing of data concerning health for the purpose of scientific research in the context of the COVID-19 outbreak*, p. 46, 21 April 2020, https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202003_healthdatascientificresearchcovid19_en.pdf [https://perma.cc/W9FZ-8FQE].

159 EDPB, *supra* note 87, at p. 31. The EDPB has noted that "anonymisation is the preferred solution as soon as the purpose of the research can be achieved without the processing of personal data."

160 Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymisation Techniques*, p. 5, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf [https://perma.cc/R9LH-78RF].

161 *Id*. at p. 8.

162 GDPR, *supra* note 6, at art. 13.

163 *Id*. at art. 9.2.j.

164 EDPB, *supra* note 158, at p. 55-56.

About the Author

Mathias Vermeulen is the public policy director at AWO, a new data rights agency, and an affiliated researcher at the Centre for Law, Science, Technology and Society at the Vrije Universiteit Brussel. Earlier, he served for four years as a digital policy advisor to Member of Parliament Marietje Schaake in the European Parliament and as the main advisor to Prof. Martin Scheinin, the U.N. special rapporteur on the protection of human rights while countering terrorism. He also advised a number of foundations, NGOs, and governments on digital rights, content regulation, and emerging technologies. Vermeulen started his professional career with internships at the Special Procedures Unit of the U.N. High Commissioner for Human Rights and at the International Commission of Jurists in Geneva. He holds a Ph.D. in European privacy and data protection law from the Vrije Universiteit Brussel.

© 2021, Mathias Vermeulen.

Acknowledgments

The author is grateful to Alex Abdo, Jef Ausloos, Cristina Blasi Casagran, Jameel Jaffer, Julian Jaursch, Amy Kapczynski, Laureline Lemoine, Paddy Leerssen, Claudia Prettner, and Rebekah Tromble for helpful guidance, feedback, and conversations on this topic and early drafts. Special thanks to Ravi Naik and Daphne Keller for providing detailed comments on earlier drafts. All errors remain my own. This research was financially supported by the Mozilla Foundation and Reset.

About the Knight First Amendment Institute

The Knight First Amendment Institute at Columbia University defends the freedoms of speech and the press in the digital age through strategic litigation, research, and public education. It promotes a system of free expression that is open and inclusive, that broadens and elevates public discourse, and that fosters creativity, accountability, and effective self-government.

knightcolumbia.org

Design: Point Five
Illustration: ©Erik Carter



KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY