Email from Paul-Olivier Dehaye, 8 March 2018

Just after writing to you on March 6th to submit evidence, I have fortuitously received a response from Facebook to two of the requests I had originally submitted to them in 2017. These were for access to personal data collected through their Custom Audiences and tracking Pixel tools. My hope is that by accessing such data, I could retroactively figure out on which webpages I was tracked, who was working with whom, etc. On top, presumably, this would open the possibility for any other Facebook user to do the same. These efforts very much echo of course the comments to your Committee by the UK Information Commissioner.

You will see below (edited only to anonymize email addresses) that Facebook's response only came after prodding by the Irish Data Protection Commissioner. Over the past year, the Irish DPC has refused to get involved, beyond encouraging Facebook to respond to me, or encouraging me to address my questions to Facebook directly.

I would recommend you read Facebook's response first. You can then read my own summary:

1) they do have the data I am asking for;
2) they retain this data "primarily for back-up purposes and data analytics";
3) they implicitly acknowledge what I am asking for indeed constitutes personal data;
4) they claim an exemption (which exists in Irish data protection law), based on "disproportionate effort", so see this purely as a technical issue;
5) they transform my request by immediately scaling expectation upwards, assuming their whole user base would make the same request;
6) they describe the difficulty of answering my request as proportional to the time scope of the request (number of hours) times the number of Facebook hourly users (since the data is not indexed by user, but retained on a per hour basis);
7) their test of proportionality is very limited, and only concerns the threats they had themselves previously identified, and particularly not against the interests I might be trying to myself preserve (e.g. enable me to understand how the advertising technology ecosystem collects data, to later try to influence me, or enable me to understand the visibility US authorities would have on my web browsing if they were to ask Facebook, particularly across platforms)

I find 2), 4), 5) and 6) taken together as particularly remarkable.

Indeed, the implications of their claims would be threefold:
- as their user base grows, they get to retain all the value from the data they collect (see 2)
- as their user base grows, their data protection obligation effectively decreases, as a result of deliberate architecture choices (see 5)
- the decrease in obligations under data protection law is proportional to the square of their user base (see 6, given 5 and the size of a daily log).

It is usually said that the dominance of an internet platform is proportional to the square of the user base (this is known as Metcalfe's Law). In other words, Facebook is claiming here that their data protection obligations is inversely proportional to their market dominance.

I hope the Committee will find this helpful, and remain at your disposal should you have any questions.

Sincerely,

Paul-Olivier Dehaye

Email from Privacy Operations Team, Facebook, to Paul-Olivier Dehaye, 7 March 2018

Dear Mr Dehaye,

We understand that you have raised concerns with the Irish Data Protection Commissioner's Office ("the DPC") regarding responses we have previously provided in relation to certain of your data access requests. In particular, we understand that you have concerns about our reliance on the exemption from data controllers' access obligations provided by section 4(9)(a) of the Irish Data Protection Acts and Recital 40 of Directive 95/46/EC, which applies in cases where the provision of data would "involve disproportionate effort"
("the exemption"). The DPC has suggested that we contact you, in order to provide more detail as to why we believe we are entitled to rely on the exemption in this case, in the hope that this will allay your concerns.

As you know, we are relying on the exemption to explain why we could not provide you with the following information you requested:

• additional details regarding the adverts you were shown as a result of advertisers' use of our Custom Audiences product; and • additional information regarding what data Facebook may have received in relation to you through the use of the Facebook pixel on third party websites.

In both cases, your requests were for information that is not available through our self-service tools which we explained to you (including DYI and Ads Preferences). Instead, this information is stored in "Hive", which is Facebook's log storage area where data is stored primarily for back-up purposes and data analytics.

Data stored in Hive is kept separate from the relational databases that power the Facebook site, and is primarily organized by hour, in log format. Hive uses this "data warehouse" architecture for resourcing reasons which are necessary in order to allow us to maintain the effective functioning of the Facebook platform, given our global user-base and the volume of data that our users create.

Importantly for current purposes, as we have sought to explain to you in previous responses, an individual's data in Hive is not readily accessible. This is because it is not stored on a per user basis. Facebook simply does not have the infrastructure capacity to store log data in Hive in a form that is indexed by user in the way that it can for production data used for the main Facebook site.

Even though we do not retain the underlying data in accessible form, to inform and provide transparency for our users we have architected systems which enable them to see the advertisers whose adverts they may be seeing because they have visited a website or app with Facebook pixels installed; as well the advertisers who adverts they may be seeing via Custom Audiences. As we have explained to you previously, this information is available in the "Advertisers whose website or app you've visited" section in Ads Preferences; and the "Advertisers with your contact info" section in Ads Preferences and DYI.

With the context explained above, when we considered your requests, we concluded the exemption would apply on the facts of this case because:

• As explained above, Hive data is not indexed by user. Instead it consists of log data stored in tables split into partitions, commonly by hour of the day.
• In order to extract user specific data from Hive, we would need to search all partitions for all possible dates in order to find any entries relating to a particular user ID.

• Running a single user query across the data stored in Hive in this way would entail multiple hours of total computing time, across thousands of servers running in parallel.
• Any access solution we provide must be available to all our users in a uniform manner – given the scale at which we operate, it is simply not feasible for us to provide solutions which are specific to individual users. Facebook receives over a million DYI requests per month and under Irish data protection law it must comply with access requests within 40 days. In that context, retrieving Hive data for all users making access requests would be technically impossible. The required computer processing power would greatly exceed that available to the Facebook group.

• The huge technical challenges presented by searching Hive for individual users' data are clearly disproportionate when balanced against the following considerations which show how any impact on our users' rights is mitigated:
o This data is also not used to directly serve the live Facebook website which users experience.
o We have nonetheless architected systems which provide our users with an intelligible representation of the data contained in Hive which you are seeking via DYI, Activity Log and Ads Preferences.
o These tools allow users to understand how their data is being used and exercise their data protection rights in a meaningful and effective way (e.g. via users' ability to see and remove advertisers named in "Advertisers with your contact info").

We hope that provides you with the information you require; however we would be very happy to answer any additional questions you may have.

Best regards,

Privacy Operations Team
Facebook

Email to Committee from Paul-Olivier Dehaye, 6 March 2018

To the Honourable Members of the DCMS Committee,

My name is Paul-Olivier Dehaye. I am a Belgian citizen living in Geneva. I am the co-founder of a startup called PersonalData.IO, aiming to help in rebuilding trust in the digital ecosystem.

I have extensive experience investigating personal data flows, which became very helpful after I read an article in December 2015 in the Guardian on the topic of Cambridge Analytica and the services offered in the Cruz campaign[1]. As a consequence, among many other actions, I have helped Carole Cadwalladr of The Observer for some of her pieces on this company, and convinced David Carroll to make his ground-breaking Subject Access Request to Cambridge Analytica[2].

[1] https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data

[2] http://motherjones.com/politics/2017/12/a-groundbreaking-case-may-force-controversial-data-firm-cambridge-analytica-to-reveal-trump-secrets/

I am writing to respectfully share with the Committee a couple elements that might help your thinking on those matters:

- what do platforms have to disclose at individual level?
- a potential problem with the new DP bill

What do platforms already have to disclose?

A lot of your Committee hearings concerned the disclosure obligations that platforms had, but were centered on top-level disclosure (to the Committee, to the regulators, etc). I actually believe that at an individual level, platforms already have a much larger obligation to disclose information, particularly under data protection laws. This is an angle that Max Schrems has used in the past, and in response Facebook was forced to adapt his systems to include a *Download your archive* tool, disclosing some of the data they held, as well as disclosing for each ad how it is targeted. Researchers have shown that for the latter Facebook is unambiguously deceptive[3], while the former is glaringly incomplete.

For instance, an individual could ask Facebook for all the advertisers who have told Facebook they had consent from that individual to add them to a Custom Audience. This might help them understand how information (false, true, misleading, inflammatory, etc) might be targeted to them, and where the advertisers were coming from.

As a test case, I did ask Facebook for this information for my own account, and again, after a long procedure through Privacy Shield, Facebook was forced to change its systems. Now *any Facebook user* who goes to "Settings → Download a copy of your data" will get that list of advertisers (try it for yourself!). In other words, my efforts have led to a direct increase in the level of disclosure through that tool.

Emboldened by this initial success and given my ongoing work with journalists, I asked in January 2017 for a lot more such data points that I knew would become relevant in an electoral context. This included for instance Facebook Pixel Data (i.e. which webpage Facebook knew I had been tracked on) or more precise information on the Custom Audience data (such as whether the connection was made through a phone number, an email, etc). My hope was that journalists could help crowdsource some efforts (even retroactive!) to understand how ads were targeted in the political arena.

It is of course extremely difficult to talk to a company like Facebook as an individual, so by April 2017 I had to escalate the matter to the Irish Data Protection Commissioner. By October 2017, after a lot of prodding, the Irish Data Protection Commissioner finally agreed to take the first step with my complaint, and ask Facebook for a comment. By December 2017, they had apparently received a response, but as of March 2018 they are still "assessing" it, despite frequent reminders. It is very hard not to see a **problem here with respect to enforcement.**

Many people say that the right of access is pointless, but I fundamentally disagree. It is true that for now it could only ever be useful to the most data-driven of activists, or to "follow the data" journalists. More or less the same is true for Freedom of Information laws, but people understand more easily the common value there. Regardless, some regulators have historically taken an

[3] See Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations, Table IV: Among all criterias used to target an ad, Facebook only discloses the most innocuous of them (for instance, if demographic and behavioural criteria were used to target an ad, Facebook would only tell the user about the demographic one). This is why, for instance,

individual's right to access their personal data in a very light-hearted way, which has compounded the problem: journalists are now discouraged from using this tool, which means we have lost a very important mechanism for accountability of the platforms. Also, journalists have not felt emboldened enough to use the courts to press for more disclosure (while they might feel inclined to do this for Freedom of Information requests).

In addition, the right to data portability (coming with GDPR) will open up new options for porting data from a platform to academic, journalistic or educational efforts, but only if the regulators encourage such reuse as much as they encourage reuse by businesses (in fact, one could argue they should encourage accountability through civil society before they encourage reuse by businesses…).

In short, I would encourage the Committee to think through at the vast possibilities for enforcement, accountability, transparency, education, scientific inquiry, journalism, etc if the right of access was made much more effective at individual level.

Problem with "Re-identification of de-identified personal data" in new DP bill

I am writing this with the clear understanding that I am not a UK citizen, but that my data is certainly processed by UK companies. The new DP bill, in its second reading, includes a provision concerning "the re-identification of de-identified personal data"[4]. In short, it limits drastically the circumstances in which someone can legally re-identify personal data that has been previously de-identified. My concern is that this provision interacts with previous jurisprudence of the European Court of Justice, in the *Patrick Breyer v Bundesrepublik Deutschland* case, in ways that are very detrimental to data subjects. In that case, the Court had to decide on whether dynamic IP addresses constituted personal data. That determination hinged on the reidentifiability of the data. The Court decided in the Breyer case that the data would not be identifiable "if the identification of the data subject was prohibited by law".

In other words, if the DP Bill passes with Provision 171 unchanged, it creates a huge loophole in the GDPR's protections. One could imagine a lot of data floating around about someone, enough for the person to be unique with those characteristics, even without any pseudonym so as not to fall within that explicit category of the GDPR (think of the game *Guess Who?,* where only Paul has glasses and white hair). Then the traceability of data for the data subjects would be broken, not in practice (like now) but really in law. The data would still single them out at any step, and retain full usefulness for its custodians at both ends of the chain, but since at least one intermediary would have no legal way to re-identify the data, suddenly that data is no longer identifiable in the sense of the CJEU, and therefore outside of the scope of protections of the GDPR. In other words, that intermediary (or a network of them) could work together to aggregate data about a white haired man with glasses from a variety of sources (insurance, web browsing, etc), without any need to worry about data protections. They could eventually pass on the result to someone else who knows about Paul and his other characteristics (white hair and glasses), and who therefore would not need themselves to perform any re-identification and risk falling afoul of Provision 171. The best way to think about these intermediaries' actions would be as "data laundering".

I hope the Committee will find my contributions useful, and remain at your disposal for any further comments or questions.

Most sincerely

Paul-Olivier Dehaye

[4] Provision 171 here: https://publications.parliament.uk/pa/bills/cbill/2017-2019/0153/18153.pdf