medConfidential comments on Bulk Personal Datasets draft codes

For paragraphs 4.17 and 4.20 of Annex A, we note that the powers that HMG wishes to give to UK agencies are why the UK research community should entirely blacklist China and other hostile states for the interactions between their researchers and their agencies. Effectively, the position of HMG appears to be that other countries should consider the UK a hostile state, with all the negative consequences for research collaborations and interaction.

Covering both Annex A & B

The codes argue that because data is held by a third party that the access should be easier. While the Agencies themselves may not care about public opinion of those powers, those who provide data may be open to wider pressures that the Codes (with an exclusive focus on Agencies) ignore.

The UK research community is increasingly moving towards making data available to users only within a "trustworthy research environment"[1] – a remote system managed by the data controller and where a researcher / user connects to perform their analyses on data rather than receiving a copy of the data under their own control – which are designed to increase the confidence of data subjects in large scale bulk personal datasets. These codes largely undermine the protections from migration to that framework, as the 3PBPD code (as well as the 7A code) both allow for agency access on the basis that when anyone has access then the agencies should have easy access.

Annex A – low or no expectation of privacy

In the event of a catastrophic data breach of vast quantities of sensitive and identifiable personal data on UK citizens being posted on the internet, the Code suggests all of that data would now be entirely fair game for the Agencies to use however they wish. Would one Agency delay taking the data down until another had been able to grab a copy under these powers? Would these powers still apply when the data had been secured? What if it was only sitting on the hard drives of a bunch of racists somewhere? Technically it would still be fair game for exploitation?

4.17:

> As written, the code allows the agencies to access and use as they see fit all the raw answers in any survey purely because the respondents to the survey consented to those running the survey doing an analysis (with the promises and stated purposes that came with it), which is blatantly untrue and entirely untenable. Any promises given on how the responses to surveys (or consultations, including this one!) run by UK companies, charities, or organisations would be entirely useless because even if the individual felt they had been given a promise of privacy by the researchers, the agencies would be able to entirely disregard it because they could get access to the data.

[1] The NHS calls these "secure data environments" and ONS calls it the "integrated data service" or "secure research service". They're all the same thing.

Surveys about sensitive patterns of behaviour would have a low level of protection in the hierarchy of data, accessible simply because an Agency thought it may be useful.

4.20:

This clause is entirely ineffective due to the overlap with “factor (d)” which implies an extraordinarily broad nature is the intent, not the narrow reading possible reading this clause alone.

There is an entire funding programme from UKRI aiming to get much more data for researchers, from smart meter data to Tesco clubcard data – the agencies saying that this data has been used for research means they can use it for their own purposes including reidentification.

Data that has been used as described includes research datasets from institutions like the UK biobank, and companies such as Facebook. Merely because research has made use of restricted access datasets on particular terms does not mean using those datasets becomes a free for all or in the “public domain” as suggested in the Code.

Given the scale of the use in “industry or academia”, it appears that the consolidated medical history of every person in England would only be excluded from this code by the nature of the fact that it is health data; ie the consolidated educational history from pre-school to post-university of every child and former-child in England since 1996 would be entirely open for use under these powers. NHS England’s claim that the data they disseminate is “anonymous”, while simultaneously noting that all their pseudonymised data with unprotected identifiers remaining remains personal data would interact in complex ways that are entirely opaque – a consequence that suggests the Code is insufficiently clear for public understanding of how powers in the Act are being used.

The current Data (Uses and Access) Bill before Parliament which enforces access to social media data for research and these powers would allow Agencies to use that for their own ends.

4.25:

This paragraph seems to give the Agencies the power to use almost any dataset it wishes for capability development, even if that dataset would be considered highly sensitive if used for operational purposes. Where real data is used for development, the standard for usage should be the same as any other use of real data – the focus is the data.

4.33:

“Very small amounts” is how much? A spoonful of sewage in a barrel of wine is “very small amounts” but still sufficient to not serve it to your friends.

Upon acquisition of a data set, the dataset should be examined with confirmation that the dataset is as expected with the obligation to notify those who authorised the use that it is as expected, or that it is not and the options for next steps. Those assessments and notifications should be included in the figures reported to IPCO and be available for inspection.

4.34:

Every event affected by 4.34 should be reported to the IPCO (allowing them to determine whether investigation is necessary). The count of events reported by each Agency should be included in the IPCO annual report.

7.16–7.19:

When discussing serious errors, the code assumes that a serious error would only ever apply to one person at a time. Given these are Bulk Personal Datasets with an assumption of no reasonable expectation of privacy, where a serious error occurs it is likely to cover many people, possibly multiple millions of people, simultaneously given some of the systems assumed by the Code and the practices of the Agencies. The process requirements of a "bulk" breach of the rights of many people should not be a barrier to determining that the error was serious and they should be informed; however making that determination one at a time may be used as a reason to not notify anyone because the Agencies are not in a position to know which of those people suffered the prejudice or harm. In a Bulk Personal Dataset (of any kind), if an error means someone has suffered prejudice or harm then everyone who may have been affected should be notified if the IPCO can not determine exactly who it was. "Following the law was too hard" should not be an acceptable justification for either Agencies or IPCO; but as written, it is.

Annex B – 3PBPD code

Annex B covers where Agencies use web interfaces (or similar, including API calls) to access the data held by other organisations.

If an Agency wants to know someone's credit history, they can use Experian's commercial search without having all the data themselves. If they want to find someone's movements or patterns of life, they can buy the most creepy tracking services they can find. New companies will be created to circumvent the few constraints that they agencies place on themselves. IPCO and oversight will have to be aware of such practices and be transparent about how they get gamed.

As part of making 3PBPD arrangements, the Code should ban Agencies from using 3PBPD services that are not a) advertised openly, and b) do not have other customers both in and outside the UK who are not public bodies. It should not be lawful to start a company that solely sells to the Agencies under 3PBPD powers to do things it would otherwise not be lawful to do under the BPD powers directly.

1.7:

The code should make clear that when datasets on multiple people (who are not likely to be the target of intelligence interest) are downloaded from the third party under this code, which they potentially can be, then they become Bulk Personal Datasets under Part 7 itself.

2.6:

The notion of “generally available” is extremely vague. The Code should clarify whether 3PBPD must follow all the same processes as the general public, or whether they can have a different process on the side of the supplier. If a supplier sees “mr.smith@gchq.gov.uk” as the login address and chooses to give full access to data that others would normally have to follow an application procedure for, would that constitute available under this Code?

Or, similar to another famous example, if the access method was “a display” in a basement of a private office where the lights had gone and so had the stairs, in a disused lavatory, behind a sign saying “beware of the privacy campaigners”, would the fact that the keyboard to type into the computer was in a locked filing cabinet make any difference to the “generally available” test? or would the filing cabinet have to be unlocked?

If Access was specifically given to the agencies by a third party that was never feasibly findable or usable by anyone other than the Agencies, would that satisfy the 3PBPD test? It should not.

2.9:

The code should clarify that making payment (or similar non-financial arrangements) is not an acceptable “specific and limited set of qualifying criteria” (or should it be sufficient, be clearer how ineffective this code is at delivering Parliament’s intent).

2.10

Related to this wider access, if an Agency is given a 99.9% discount on list price which is set so high no one purchases it, would that be considered “generally available”? Would the same be true of a 90% discount? Or a 10% discount?

3.10 footnote 5

The use of the word “obtained” interacts very weirdly with the notion of 3PBPD access. Only the search results copied into possession of the Agencies could be said to be “obtained” by the Agency, despite having run searches on swathes of data which the Code is supposed to regulate.

6.17-6.20

Our comments on serious harm made in 7.16-7.19 of Annex A apply, but note that the harm may be caused by the Third Party provider. Mistakes by that provider, especially relating to data held by commercial entities, may be entirely impossible to determine without the assistance from the commercial entity which they would be under no obligation to provide.

The most common example of failing in this category would be the annual revelation of mistakes covered in IPCO annual reports where a request for communications data is made to a communications service provider and something gets garbled and information on someone else is provided. While those requests are lookups for communications data not lookups for 3PBPD, that there is a steady series of new failure modes and utter cockups in a system that is recognised as being highly sensitive and of utmost importance shows that lookups through 3PBPD are not infallible, and indeed should be expected to regularly go wrong. The suggestion that mistakes will be one offs and rare is not borne out by the IPCO reports.

The Code is also unclear how the statutory standards required by law, and the IPCO investigatory processes required, can ever be met under the “generally available” standard – that agreement will always require additional access as defined in 2.10.

We assume that Agencies would prefer to have control of key BPDs for their own purposes, and will only use 3PBPDs when they are necessary and bring data within the Part 7 Code as soon as possible. However, should practise diverge from that assumption over time, there may have to be additional IPCO scrutiny and a Code revision. There will be entities who spring up to service the giant gaps in this code under contract that are nominally available to anyone, but which are only in practice sold to Agencies, allies, and those who pretend to be allies (as the companies serving the US government keep finding out by accident).

Missing:

Given this Code is likely to serve as the only public reference point of how these powers are used, the Code should also be very clear that no third party is ever obliged to give 3PBPD access if they do not wish to (ie they reserve the right to refuse access to anyone else), or the agencies do not meet their published (or unpublished) criteria. The owners and controllers of many datasets have particular criteria for third party access that should be respected at all times by the Agencies using powers under the 3PBPD Code.

While the content of the Data (Uses and Access) Bill is currently a matter for Parliament, the Code will need to be amended when what was initially Schedule 4 Clause 2 has passed. The consequence of these two powers taken together is that every dataset analysed within the UK jurisdiction would be available under 3PBPD powers if it is available to any researcher. This creates a chilling effect for others to give access to UK based researchers while taking out billboard ads for the most creepy and intrusive to offer services.

medConfidential
November 2024