Is It Legal? Is It Right? The Can and Should of Use

A t three years after Edward Snowden’s disclosures of National Security Agency (NSA) surveillance, it’s a good time to reflect on lessons learned. The issues that raised the most concern included the bulk collection of domestic metadata, authorized through a secret interpretation of a US law (Section 215 of the USA PATRIOT Act), and the warrantless collection of communications where one end was outside the US, which was explicitly authorized under Section 702 of the FISA (Foreign Intelligence Surveillance Act) Amendments Act. Meanwhile, US companies were incensed over NSA’s purported hacking into the communications of its overseas datacenters (domestic interception would have required a wiretap warrant). Moreover, NSA’s subversion of the National Institute of Standards and Technology’s (NIST’s) process for recommending cryptographic algorithms undermined NIST’s ability to effectively develop cryptographic and security standards, which could severely impact national and global cybersecurity. There was also indignation over the US tapping the communications of leaders of close allies, including German Chancellor Angela Merkel.

Given the depth of the anger, it might be somewhat surprising that, with the exception of the bulk collection of metadata, there was little disagreement over the legality of NSA’s data collection. The warrantless collection of communications with at least one foreign component had been publicly debated before being made into law; from the point of view of US law, the capture wasn’t illegal. Although NSA’s action caused NIST to recommend a presumably hacked algorithm as a pseudorandom-bit generator, the action wasn’t illegal. Wiretapping the German chancellor violated German law, but much foreign intelligence gathering violates the laws of the nation in which it takes place. It’s a truth universally acknowledged that nations have intelligence agencies whose job it is to spy, even on allies.

The furor and response to the Snowden disclosures make clear that following the law, while necessary, wasn’t sufficient. Despite the apparent legality of the NSA surveillance, Snowden’s disclosures led to changes in laws and policies. Being legal isn’t enough; other issues must be addressed as well.

University of Michigan law professor Margo Schlanger observed that, prior to the Snowden disclosures, the NSA and the administration’s focus was on “can we (legally) do X?,” and not “should we do X?”1 Such “intelligence legalism” has three features: substantive rules that are treated as laws, limited court enforcement of those rules; and empowerment of lawyers.1 Although many smart people thought long and hard about the various forms of surveillance—and there were numerous disputes within the government about it2—the questions of “is it just?,” “is it moral?,” and “is it appropriate?” don’t appear to have had much debate. It was only after the Snowden disclosures that these harder questions were asked—or at least asked loudly enough that change resulted.

The surveillance didn’t stand up to public scrutiny. The Merkel wiretapping didn’t pass the front-page-of-the-Washington-Post test (“Yes, it’s legal, but do you want this story on the front page of X newspaper?”). The Privacy and Civil Liberties Board’s case-by-case study of the bulk domestic metadata collection showed that the data’s value in terrorism cases had been minimal.3 (One minor domestic case was uncovered as a result of the surveillance,3 and communications metadata had been used to rule out foreign terrorists having a US connection.4) A result of all the scrutiny was the 2015 USA Freedom Act; this ended the bulk collection of domestic telephone metadata by NSA. Such data, when held at the phone companies, could be searched by NSA under court order.

The public discussion that occurred after the Snowden disclosures shows that
government decisions regarding surveillance shouldn’t be based solely on whether the decisions are legal but also on whether they’re right. The question of right encompasses many more issues, such as the value of the surveillance and cost to intangibles including both real and perceived threats to privacy. “Should we?” is far more complex to resolve than “is it legal?”

The Government Isn’t the Only One Conducting Surveillance

Governments aren’t the only organizations collecting massive amounts of data on private individuals. Indeed, it might well be the case that the government’s data collection on individuals is dwarfed by the private sector’s.

There’s much to be gained by private-sector data collection. Google’s search responses are tuned by information gleaned from other people’s searches and responses, while its maps application can provide real-time navigation advice based on the information from other app users’ simultaneous use. Medical and computer science researchers discovered they could anticipate queries concerning diagnosis of pancreatic cancer based on earlier Bing search queries. Pancreatic cancer is very hard to treat because it presents late—so this work is intriguing. Could one advise that people go to the doctor based on their queries? Could this lead to early diagnosis? Early treatment? Whether it’s better scheduling of rush-hour transit, understanding the flow of people during a natural disaster, tracking the path of a disease, or updating how information is presented on smartphone screens, it’s clear that big data provides value.

The value of the surveillance’s result doesn’t allow dodging the “can we?”/“should we” issue. The President’s Advisory Council on Science and Technology report, Big Data and Privacy: A Technological Perspective, argues that big data is so useful that it will be collected. Rather than controlling collection, privacy preservation must focus on controlling the data’s use. This makes sense. As a small example, consider the fact that smartphone OS providers capture information about the order in which users conduct swipes on the phones. Studying this metadata can help improve the product’s design. But this same data could also be used in other ways, ones that might not please the user, for instance, to infer mood and control access to certain devices as a result. The former passes the “should we” test—virtually every user would find such use beneficial— whereas the latter does not.

Consider the pancreatic cancer study with the “can we?”/“should we” issue in mind. The researchers did something very useful: they found that patients who don’t yet present symptoms of pancreatic cancer might nonetheless have hints of their illness hidden in their searches, hints that reveal the disease before it might be discovered otherwise. But reflect on how the study was conducted. The researchers explained:

Every interaction corresponds to a log entry containing the query, the results selected, and a timestamp. A unique, anonymized identifier linked to the Web browser is also included, enabling the extraction of search log histories for up to 18 months. The anonymous identifier is tied to a single machine. … We used proprietary logs from Bing .com for searchers in the English-speaking United States locale, from October 2013 to May 2015 (inclusive).

The Microsoft privacy notice included a statement about performing research on submitted data. But I doubt Bing users anticipated that their search terms would be studied to determine whether a pancreatic cancer diagnosis lay in their future.

Ethical Choices

Let me clarify the distinctions between this study and some earlier research conducted on users without their explicit opt-in. I’m thinking here, of course, of the Facebook study that manipulated the news feeds of 70,000 users to examine their reactions when served “happy” and “sad” posts. (Facebook viewed its terms of service, which didn’t include discussion of use for research, as providing informed consent; the company later updated its terms of service to include the fact that it conducts research and tests features on its users [www.facebook .com/about/privacy/update#].) The pancreatic cancer study is quite different: it examined query data and its link to a diagnosis of pancreatic cancer. It didn’t manipulate users; it simply studied the search data and found some potentially interesting correlations.

I nonetheless find the study’s use of search data “creepy.” The pancreatic cancer study uncovers very personal information about individuals, information they might not want others to know—indeed, information they might not want to know themselves. I might be
willing to participate in a study that examines my searches for potential diseases, but my willingness would depend on which diseases are being studied, how and with whom the data is shared, how the data is used, and so on. Without question, though, I would want to be explicitly asked whether I’m willing to volunteer for such a study.

It does not appear that the users in this particular study were asked about their participation. Rather, researchers used the Bing database of user searches to conduct the study, with the idea that anonymizing the identifier would provide sufficient protection. But the identifier was tied to a single machine, and in such a case, there’s no anonymity. More than a decade ago, we learned that multiple query searches can easily be used to reidentify so-called “anonymized” users. If a search engine’s policy indicates it can use my data for research of such a personal nature without first checking with me for approval, I’ll take my searches elsewhere, such as DuckDuckGo, which protects user privacy, or use a search engine through Tor, which prevents tracking of my machine.

In the wake of the Snowden disclosures, the US government learned that the “should we?” question is as important to ask as the “can we (legally)?” question. Big data has much to offer society, both in the aggregate—for instance, transit planning—and at the individual level—for instance, the pancreatic cancer study. However, as with government surveillance, private sector use of big data necessitates asking “should we?” in addition to “can we (legally)?” One of the authors of the pancreatic study wrote, “Improving the transparency of data processing to data subjects is both important and challenging.” I would put it differently: improving the transparency, while challenging, is absolutely necessary.

“Should we?” questions include: Is this an application of the data for which a significant percentage of users might hesitate to participate if asked? Is this use of data substantively different in nature from “normal” usage? Might the user feel manipulated or intruded on as a result of this use of the data? If we are using people’s private data to inform our results, we have a moral and ethical responsibility to ask such questions. Anything less fails our users.

References