Daniel E. Geer Jr.
In-Q-Tel

Personal Data and Government Surveillance

In September 2013, the US National Academy of Sciences, on behalf of the Department of Homeland Security, concluded that cybersecurity should be seen as an occupation and not a profession because the rate of change is too great to consider professionalization (www.nap.edu/openbook.php?record_id=18446). That rate of change is why cybersecurity is perhaps the most intellectually demanding occupation on the planet. The fundamental intellectual challenge might be understood by narrowing our focus to one class of tradeoffs in cybersecurity: personal data and the government.

Data are what cybersecurity is all about. In 2007, Jim Gray gave a seminal talk about key transformations in the history of science, coining the term "fourth paradigm" (http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt). By that, he meant that science began as an endeavor organized around empirical observation. Next came the age of theory: theorizing as the paradigm of what science did. Then science became computational, meaning that the paradigm of science was to calculate. Gray argued that we're now in a fourth era that has shifted from computational science to data-intensive science.

To see how, consider ecology professor Philip Greear: he would challenge his graduate students to catalog all the life in a cubic yard of forest floor. Similarly, computer science professor Donald Knuth would challenge his graduate students to catalog everything their computers had done in the last 10 seconds. It's hard to say which is more difficult, but everywhere you look, cybersecurity practitioners are trying to get a handle on "What is normal?" so that that which is abnormal can be identified early in the game. Behavioral approaches leading to intrusion detection are exactly the search for anomaly, and they're data based. The now-famous attack on RSA Data Security that led to RSA's buying NetWitness is an example of wanting to know everything so as to recognize something. The central organizing principle behind a competent security program is to instrument your data sufficiently well that nothing moves without being noticed. While physics has made it possible to put computers everywhere, it has also made it possible to fill them all with data. Cybersecurity is barely keeping up.

Twenty Questions

Imagine playing a game in public where I ask you a mildly embarrassing question, something as mild as, "How many pairs of never-used underwear do you own?" Then a second question, something similarly mild, such as, "Have you ever had an evil grin while wrapping a birthday gift?" Most will become uneasy and few indeed will get to the proverbial "20 questions." Why? Because the subject realizes that data fusion of even mild, innocuous questions has the effect of painting a picture. In fact, the more inane the questions are, the more inane the picture painted becomes.

If you get to pick the questions and your subject is willing to keep answering them, then you can pretty much box in your subject however you like. Politicians know that the surest way to win an argument is to, as they say, "frame the question," by which they mean painting a picture that their opposition has to work to overcome. The better practitioners at the political version of this game can impose a considerable work factor on their opponents, one that is not unlike what we call a denial of service in the computer realm.

When I worked for a data protection company, our product was, and I believe still is, the most thorough on the market. By "thorough" I mean the dictionary definition: "careful about doing something in an accurate and exact way." To this end, installing our product instrumented every system call on the target machine. Data didn't and couldn't move in any

1540-7993/14/$31.00 © 2014 IEEE

sense of the word without detection. Every data operation was caught and monitored. The company's customers don't accept half-measures. What made this product stick out was its very thoroughness, but here's the point: unless you fully instrument your data handling, it isn't possible for you to say what didn't happen. With total surveillance, and total surveillance alone, it's possible to treat the absence of evidence as the evidence of absence. Only when you know everything that *did* happen with your data can you say what did *not* happen with your data.

The alternative to total surveillance of data handling is to answer narrower questions, such as, "Can the user steal data with a USB stick?" or "Does this outbound email have a Social Security number in it?" Answering direct questions is exactly what a defensive mindset says you must do, which is "never make the same mistake twice." In other words, if someone has lost data because of misuse of some facility on the computer, then you either disable that facility or you wrap it in some kind of perimeter. Lather, rinse, and repeat. This extends all the way to such trivial matters as timer-based screen locking.

Shifting Mindsets

The difficulty with the defensive mindset is that it leaves in place the fundamental strategic asymmetry of cybersecurity—namely, that while the work factor for the offender is the price of finding a new method of attack, the work factor for the defender is the cumulative cost of forever defending against all attack methods yet discovered. Over time, the curve for the cost of finding a new attack and the curve for the cost of defending against all attacks to date cross. Once that happens, the offender never has to worry about being out of money. I believe that this crossing occurred some time ago.

The total surveillance strategy is, to my mind, an offensive strategy used for defensive purposes. It says, "I don't know what the opposition is going to try, so everything is forbidden unless we know it's good." In that sense, it's like whitelisting applications. Taking either the application whitelisting or the total data surveillance approach is saying, "That which is not permitted is forbidden."

But the essential character of a free society is the opposite: that which is not forbidden is permitted. The US began as a free society without question; the weight of regulation, whether open or implicit, can only push it toward being unfree. Under the pressure to defend against offenders with a permanent structural advantage, defenders who opt for forbidding anything that isn't expressly permitted are encouraging a computing environment that doesn't embody the freedom with which we are heretofore familiar.

We know that more and more data are in play, more and more data are collected. The general dynamics of change are these: Moore's law has given us two orders of magnitude in compute power per dollar per decade, while storage has grown at three orders of magnitude and bandwidth at four. These are top-down economic drivers. If these continue, the future can only be increasingly dense with stored data, but, paradoxically, despite the massive growth of data volume, those data would become more mobile with time.

One of the defender's highest goals is to minimize the attack surface wherever possible. Every coder adhering to a security-cognizant software life-cycle program does this. Every company or research group engaged in static analysis of binaries does this. Every agency enforcing a need-to-know regime for data access does this. Every individual who reserves one low-limit credit card for Internet purchases does this. I might otherwise say that any person who encrypts email to his or her closest counterparties does this, but because consistent email encryption is so rare, encrypting your email marks it for collection and indefinite retention by those entities in a position to do so, regardless of what country you live in.

Data retention for observable data is growing by legislative fiat seemingly everywhere. The narrow logic is sound—namely, if data have passed through your hands, then your retention of them has no new risk for the transmitter and might offer valuable protections against malfeasance. More broadly, neither you nor I would be concerned with some entity having access to one of our transmitted messages, but 1,000 of them is a different story, and all of them forever is a different world.

Digital Conscripts

In the US, almost all the critical infrastructure is in private hands, and Internet-dependent services are becoming more essential to "normal life." The government's response to the growing pervasiveness of Internet services held in private hands is to deputize the owners of those services against their will to capture and provide all kinds of data. That is, if the government does not itself own the critical infrastructure, those who do can and will be compelled to become government agents. Ironically, we have a physical army of volunteers but a digital army of conscripts.

Data are at the core of it all. The great majority of attacks target data acquisition. Indeed, the work of surveillance is targeted data acquisition. Yet the Federal Communications Commission classifies the Internet as an information service, not a communications service. Although that might have been a gambit to relieve ISPs of telephone-era regulation, the value of the Internet is ever more the bits it carries, not the carriage of those bits. The FCC has made decisions that are both several and old by classifying

- cable as an information service in 2002,
- DSL as an information service in 2005,
- wireless broadband as an information service in 2007, and
- broadband over power lines as an information service in 2008.

A decision by the DC Circuit Court of Appeals on this very last point is pending: Is the Internet a telecommunications service or an information service? The difference is crucial. If an ISP is an information service, it can charge whatever it likes based on the contents of what it's carrying, but it's responsible for that content if it is illegal. Inspecting brings with it a responsibility for what it learns. If an ISP is a telecommunications service, it can enjoy common carrier protections at all times, but it can neither inspect nor act on the contents of what it carries and can only charge for carriage itself; in other words, bits are bits.

Flash Crash

We humans can design systems more complex than we can then operate. The financial sector's "flash crashes" are a good example; perhaps the 50 interlocked insurance exchanges that comprise the Affordable Care Act's mandate will soon be another. Above some threshold of system complexity, it is no longer possible to test; it is possible only to react to emergent behavior. Even the lowliest Internet user is involved in the complexity: one webpage can easily touch scores of different domains. Sampling the top-level page from cnn.com showed 400 out-references to 85 unique domains, each of which is likely to be similarly constructed and all of which move data one way or another. If you leave these pages up and they autorefresh, moving to a new network signals your move to every one of those advertising networks.

When we ask the government to provide services that can be done only with more data, we're asking government to collect more data.

We know that traffic analysis is more powerful than content analysis; if I know everything about to whom you communicate, including when, where, with what inter-message latency, by what protocol, from what addresses, and at what length, then I know you. If all I have is the undated, unaddressed text of your messages, then I'm an archaeologist, not a case officer. The soothing mendacity of saying, "It's only metadata" relies on the listener's ignorance.

Governments need intelligence, and the intelligence community operates under the rules it knows, trying to reach the goal states it has been tasked to achieve. Thus, the center of gravity for policy is those goal states.

Knowledge is power, and there is a subtle yet important distinction between information and knowledge. We all know that proving a negative requires omniscience. At the same time, the more technological the society becomes, the greater the dynamic range of possible failures. For a cave-dweller, starvation, predators, disease, and lightning represent the full range of failures that end life. For a member of a technological society, where everybody and everything are optimized in some way akin to just-in-time delivery, the dynamic range of failures is incomprehensibly larger and largely incomprehensible. The wider the dynamic range of failure, the more prevention is the watchword. Cadres of people charged with defending masses of other people must focus on prevention, and prevention is all about proving negatives. Therefore, as our technological society grows more interdependent within itself, the more it must rely on prediction based on data collected in broad, not targeted, ways.

To this end, intelligence agencies that collect up everything are reacting rationally to the demand that they ensure "never" comes true. Moreover, the more complex the society they're charged with protecting becomes, the more they must surveil, the more they must analyze, the more data fusion becomes their primary focus.

Greater Than the Parts

Technology is today far more democratically available than it was yesterday and less than it will be tomorrow. 3D printing, the whole "maker" community, DIY biology, microdrones, search, and constant contact with whomever you choose—all are democratizing technologies. But they represent our last fundamental tradeoff: Do we, as a society, want the comfort and convenience of increasingly invisible digital integration enough to pay for those benefits with the liberties that must be given up to be protected from the downsides of that integration?

We must make this choice while choice is still relevant. In our service economy, every time an existing service consolidates into the cloud, our vulnerability to its absence increases. For example,

- yesterday we asked, "How do you feel about traffic jam detection based on the handoff rate between cell towers of those cell phones in use in cars on the road?";

- today we ask, "How do you feel about auto insurance prices informed by a daily readout of your automobile's black box?"; and
- tomorrow we may ask, "In what calendar year will compulsory auto insurance be more expensive for the driver who insists on driving a car himself or herself, rather than letting a robot do it? How do you feel about public health surveillance done by requiring search engine providers to report on searches for health-related topics? Would you install a toilet that does urinalysis with every use? How do you feel about a smart grid that reduces your power costs and greens the atmosphere but reports minute-by-minute about what's on and off in your home?

At this moment in time, facial recognition is possible at 500 meters, iris recognition is possible at 50 meters, and heart-beat recognition is possible at 5 meters. Your dog can identify you by smell; so, too, can an electronic dog's nose. Your mobile phone's accelerometer is sensitive enough to identify you by gait analysis. More than 3 billion new photos are placed online each month—even if you've never uploaded photos of yourself, someone else has. All of these techniques are data dependent, cheap, and convenient, and none reveals anything secret. Yet the sum of them is greater than the parts. How do you feel about using standoff biometrics as a solution to authentication?

Before you express discomfort, remember the reasons that passwords are a problem. At the same time, passwords may be essential (www.wired.com/opinion/2013/09/the-unexpected-result-of-fingerprint-authentication-that-you-cant-take-the-fifth):

> If the police try to force you to divulge the combination to a wall safe, your response would reveal the contents of your mind and so would implicate the Fifth Amendment. (If you've written down the combination on a piece of paper and the police demand that you give it to them, that may be a different story.)
>
> To invoke Fifth Amendment protection, there may be a difference between things we have or are—and things we know. The important feature about PINs and passwords is that they're generally something we know. These memory-based authenticators are the type of fact that benefit[s] from strong Fifth Amendment protection should the government try to make us turn them over against our will. Indeed, last year a federal appeals court held that a man could not be forced by the government to decrypt data.
>
> But if we move toward authentication systems based solely on physical tokens or biometrics—things we have or things we are, rather than things we remember—the government could demand that we produce them without implicating anything we know. Which would make it less likely that a valid privilege against self-incrimination would apply.

A court could find otherwise and set a different precedent, but Marcia Hofmann's analysis is cautionary. Perhaps a balance of power requires an individual to retain some secrets. But is having some secrets the same as having some privacy?

Observability

No society and no people need rules against things that are impossible. If I observe a couple fornicating in circumstances where I can never know who they are, does the couple have privacy? The answer is "no" if your definition of privacy is the absence of observability. The answer is "yes" if your definition of privacy is the absence of identifiability.

Technical progress in image acquisition guarantees observability everywhere now, and those standoff biometrics are delivering multifactor identifiability at ever-greater distances. We may soon live in a society where identity isn't an assertion like, "My name is Dan," but rather an observation like, "Sensors confirm that is Dan." How many sensors are we installing in normal life? How many are we installing in our prisons?

If data kill both privacy as impossible to observe and privacy as impossible to identify, then what might be an alternative? If you're an optimist or an apparatchik, then your answer will tend toward rules of procedure administered by a government you trust or control. If you're a pessimist or a hacker/maker, then your answer will tend toward the operational, and your definition of a state of privacy will match mine: the effective capacity to misrepresent yourself.

Misrepresentation is using disinformation to frustrate the data fusion of whoever is watching you, such as

- paying your therapist in cash under an assumed name,
- swapping affinity cards at random with like-minded folks,
- arming yourself not at Walmart but in living rooms,
- keeping an inventory of misconfigured webservers to proxy through,
- putting a motor-generator between you and the smart grid,
- using Tor for no reason at all,
- hiding in plain sight when there's nowhere else to hide, and
- having as many digital identities as you can.

Your identity is not a question unless you work to make it be.

The US National Strategy for Trusted Identities in Cyberspace "calls for the development of interoperable technology standards and policies—an 'Identity Ecosystem'—where individuals, organizations, and underlying infrastructure—such as routers and servers—can be authoritatively authenticated" (www.whitehouse.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf). The basic premise is that you could trust such a digital identity because it couldn't be faked. The government cares because it wants to digitally deliver government services and it wants attribution. Is having a non-fake-able digital identity for government services worth the registration of your remaining secrets with that government? Is there any real difference between a system that permits easy, secure, identity-based services and a surveillance system? Do you trust those who hold surveillance data on you over the long haul—that is, the indefinite retention of transactional data between government services and you, the individual required to proffer a non-fake-able identity to engage in those transactions? If you're building authentication systems today, then you're playing in this league.

Engineers often convey system choices as "fast, cheap, reliable: choose two." For cybersecurity policy makers, it's "freedom, security, convenience: choose two."

Standoff biometry by itself terminates the argument over whether security-privacy tradeoffs are a zero sum game. The sum is nowhere near that good, and it's the surveilled who are capitalizing the system. Entirely innocuous things become problematic when surveilled. Shoshana Zuboff called this "anticipatory conformity" and said (www.faz.net/aktuell/feuilleton/the-surveillance-paradigm-be-the-friction-our-response-to-the-new-lords-of-the-ring-12241996.html),

> [W]e anticipate surveillance and we conform, and we do that with awareness. We know, for example, when we're going through the security line at the airport not to make jokes about terrorists or we'll get nailed, and nobody wants to get nailed for cracking a joke. It's within our awareness to self-censor. And that self-censorship represents a diminution of our freedom. We self-censor not only to follow the rules, but also to avoid the shame of being publicly singled out. Once anticipatory conformity becomes second nature, it becomes progressively easier for people to adapt to new impositions on their privacy, their freedoms. The habit has been set.

Leonard Downie, former *Washington Post* executive editor, wrote (www.washingtonpost.com/opinions/in-obamas-war-on-leaks-reporters-fight-back/2013/10/04/70231e1c-2aeb-11e3-b139-029811dbb57f_print.html),

> Many reporters covering national security and government policy in Washington these days are taking precautions to keep their sources from becoming casualties in the Obama administration's war on leaks. They and their remaining government sources often avoid telephone conversations and email exchanges, arranging furtive one-on-one meetings instead. A few news organizations have even set up separate computer networks and safe rooms for journalists trained in encryption and other ways to thwart surveillance.

The Gordian Knot

Engineers often convey system choices as "fast, cheap, reliable: choose two." For cybersecurity policy makers, it's "freedom, security, convenience: choose two." But the choice must be made by the public at large, not by those who are trying to deliver failure-proof protection to an impatient, risk-averse, gadget-addicted population.

In the financial crisis, we saw that levels of achievable financial return require levels of unsustainable financial risk. That lesson was learned on the large scale and the small, on the national scale and on the personal one. Let's not have to learn the parallel lesson with respect to data that power the good versus data that power the bad. Think of data as a kind of money: investing too much of our own data in an institution too big to influence is just as insensate as investing too much of our own money in an institution too big to fail.

All security tools and all the data they acquire are dual use: the security tools and their data can be used for good or for ill. The wellspring of risk is dependence, especially on expectations of system state. If you're most at risk from the things on which you most depend, then damping dependence is the cheapest, most straightforward, lowest latency way to damp risk. John Gilmore famously said, "Never give a government a power you wouldn't want a despot to have." My rewrite would be: "Never demand the government have a power you wouldn't want a despot to have."

A state of security is one with no unmitigatable surprise, that is, when you can mitigate the surprises you will face. California Senate Bill 1386, the first of the state-level data breach laws, didn't criminalize losing credit card data; rather, it prescribed the actions that a firm must take when it has lost its customers' credit card data. SB1386 is wise in that regard.

But only rarely do we ask our legislatures to make mitigation effective. Instead, we repeatedly ask them to make failure impossible. But doing so forces us into cost-benefit analyses where at least one of the variables is infinite. It isn't heartless to say that if every human life is actually priceless, then it follows that there will never be enough money. Similarly, it's not anti-government to say that doing a good job at preventing terrorism is safer than doing a perfect job.

So here's the Gordian knot: as society becomes more technological, even the mundane comes to depend on distant digital perfection. For instance, our food pipeline contains less than a week's supply. It depends on digital services for everything from GPS-driven tractors to robot vegetable-sorting machinery to irrigation-monitoring drones to coast-to-coast logistics to RFID-tagged livestock. Are the technological dependency and the data that fuel it making us more resilient or more fragile?

In cybersecurity practice, we seem to be getting better and better. We have better tools, better understood practices, and more colleagues. But in the ratio of skill to challenge, we're expanding the society-wide attack surface faster than our collection of tools, practices, and colleagues. If you're growing more food, that's great, but if your population is growing faster than your improvements in food production can keep up, that's bad. As with most decision-making under uncertainty, statistics have a role, particularly ratio statistics that magnify trends so that the latency of feedback from policy changes is more quickly clear. Yet statistics, too, require data.

Privacy

We have well-established and helpful rules about medical privacy. But those rules also have huge holes. When you check into the hospital, there's an accountability-based, need-to-know regime that governs your data most days. However, if you check in with bubonic plague or anthrax, you'll have zero privacy as those are mandatory data-reporting conditions. Would it make similar sense for the public health of the Internet to have a mandatory reporting regime for cybersecurity failures? Do you favor having to report penetrations of your firm or household to the government or face criminal charges for failing to make that report? Are those data that you want to share? Sharing them can only harm you, but they might help others.

This isn't about you personally. It's about a culture where personal data are increasingly public data, assembled en masse. In the US, all we have to go on now is the hopeful phrase, "a reasonable expectation of privacy." But what's reasonable when one-inch block letters can be read from orbit, and when all financial or medical records are digitized and available primarily over the Internet? Do you want ISPs to retain emails when you ask your doctor a medical question (or, for that matter, do you want those emails to become part of your electronic health record)? Who owns your medical data, anyway? In the US, until the 1970s, it was the patient, but subsequent regulations have made it the provider. With an EHR, it's likely eventually to revert to patient ownership. But if the EHR belongs to you, can you surveil its use by medical providers and those to whom they outsource? If not, why not?

Observability is fast extending to devices, and some of it has already appeared. For instance, any newish car is broadcasting four unique Bluetooth radio IDs, one for each tire's valve stem. We train our youngsters to accept surveillance by stuffing a locator beacon in their backpacks as soon as they go off to kindergarten. We're now surrounded by cameras. A single camera might not seem important, but cameras are important in the aggregate when their data are fused. And anything with "wireless" in its name creates an opportunity for traffic analysis.

The days of radio brought Sarnoff's law: the value of a broadcast network was proportional to N, the number of listeners. Then came packetized network communications and Metcalfe's law: the value of a network was proportional to N squared, the number of possible two-way conversations. Now we have Reed's law, where a network's value is proportional to the number of groups that can form in it: 2^N. Reed's law reflects the new reality in the age of social networks. Because everything is dual use, any entity (such as a government) that can acquire the entirety of all social media transactions learns nearly everything there is to learn, all in one place, courtesy of the participants themselves. The growth of social networks is a surveiller's dream come true.

Similarly, from a security person's point of view, total system complexity is just geometry. Security is noncomposable: we can get insecure results even when our systems are assembled from secure components. And the more components, the less likely we are to get a secure result. Might the same be said of data? Of course! Search for the term "reidentification," and you'll find that incomplete data, even intentionally anonymized data, can be put together again if there's enough of it (and what's enough seems to be a lower hurdle every year). Put differently, if you share a different fact each with 10 different people, how many of the 10 have to be compromised before you're exposed?

David Brin was the first to suggest that if you lose control over what's collected about you, the only freedom-preserving alternative is to allow everyone else to do it, too (*The Transparent Society*, Perseus, 1998). If the government or the corporation can surveil you without asking, then the balance of power is preserved when you can surveil them without asking. Bruce Schneier countered that preserving the balance of power doesn't mean much if the effect of new information is nonlinear: if new information is the exponent in an equation, not one more factor in a linear sum. Resolving that debate requires having a strong opinion about what data fusion means operationally to you, to others, and to society.

There is some axiom of nature at work here. Decision-making under uncertainty is what we do in the small and what policy makers do in the large. Uncertainty reflects having only partial information, so it's natural to want information that's less partial. We're closing in on having more information than we can use. The intelligence community has felt the heat of too much information to handle for some time. The business community is feeling it now, because it's far cheaper to keep everything than to do careful, selective deletion. The individual is feeling pretty warm, too, as evidenced by increasing dependence on the ability to search email again, rather than storing it in an organized way after reading it. And then there's the political sphere (www.economist.com/news/united-states/21601516-politicians-have-never-had-access-so-much-data-how-come-their-debates-are-so).

I am old enough that I can opt out of many corporate data collection schemes and live out the remainder of my days unaffected by what I might be missing out on. That those corporations are agents of government data collection means that for now I'm opting out of some of that as well. This generational divide is leading to a kind of structural polarization. Confirm this by asking the best cybersecurity people what they will and won't do on the Internet. You'll find the responses to be sharply different from what the public at large does and doesn't do. The best people know the most, and they're withdrawing, rejecting technologies. To use the words and style of the intelligence community, they're compartmentalizing.

Anyone under 40 has no such option, or at least no such easy option. To you I say that it's your responsibility to choose whether to demand protections, conveniences, and services that can be done only with pervasive data. It's your responsibility to choose whether to fear only fear itself or to fear the absence of fear. It's your responsibility to choose whether to be part of the problem or part of the solution.

How do you make that choice? Any finite tolerance for risk caps the amount of information you will want in play, a tolerance that has nothing to do with whether you have anything to hide. It's your responsibility to choose whether to make it understood that just as "there is nothing sinister in so arranging one's affairs as to [minimize] taxes" neither is there anything sinister in minimizing the data collectible from you (http://law.justia.com/cases/federal/appellate-courts/F2/159/848/1565902/). The price of freedom is the probability of crime. But as technology progresses, your choice will not be between Big Brother and no Big Brother. Rather, it's already between one Big Brother and a lot of Little Brothers. Think carefully. Yours is the last generation that will have a choice. ■

Daniel E. Geer Jr. is CISO for In-Q-Tel and past president of the Usenix Association. Contact him at dan@geer.org.