The Code of Privacy[1]

LAWRENCE LESSIG
C. Wendell and Edith M. Carlsmith Professor of Law
Executive Director, Center for Internet and Society
Stanford Law School

IN AN ARTICLE published in 2000, Harvard Law Professor Jonathan Zittrain observed that the problem of privacy on the Internet was the same as the problem of copyright.[2] In both cases, Zittrain argued, someone's "data" had gotten out of control. In the case of copyright, the data are the copyright owner's music ripped into mp3s and spread across the Internet billions of times each day. In the case of privacy, the data are data about us increasingly gathered by the gaggles of technologies designed to aggregate data and do stuff with it. In both cases, technology had set data free from their original—and here's the assumption—proper—controller. In both cases, the problem for policy makers then is how best to return control to the data owners.

Zittrain's argument is important and right. In this brief intervention, I want to extend it as it affects privacy on the Internet. For the two "problems" of copyright and privacy could be solved not just by returning control to the proper controller. The two problems could also be solved by a regime that produced the values that copyright and privacy seek without giving anyone "control" over the data they affect.

This is a solution that is increasingly familiar in the context of copyright. Copyright law is designed to give authors exclusive rights over copyrighted material. Those rights have traditionally meant control over the ability to copy or distribute copyrighted material. The Internet has made that type of control especially difficult. The essence of this network is the capacity to copy. Copying—for a digital network—is like breathing is to us.

So many have begun to describe alternatives to the present copyright system—a regime that would not give authors exclusive control over copies, but that still gives authors the incentives to produce. A

[1] Read 24 April 2004, as part of the symposium "Privacy."

[2] Jonathan Zittrain, "What the Publisher Can Teach the Patient: Intellectual Property and Privacy in an Era of Trusted Privication," *52 Stan. L. Rev.* 1201 (2000).

compulsory license, for example, tracking use rather than controlling access, is one such example of a copyright regime that does not rely upon controlling copies. Those alternatives preserve the objective of copyright—that authors get paid—without destroying the essence of the network—that copies as it breathes.

The same should be considered in the context of privacy. Here, too, the Internet has made control over data especially difficult. Here, too, we should seek alternatives to a regime that blocks access to data, alternatives that give individuals the values that privacy seeks to assure. Is it possible to imagine a world that protected "privacy" without systematically blocking access to data? Could we achieve "privacy" without keeping things "private"?

The first instinct for those of us who believe strongly in the values of privacy is to say no: privacy can only be protected if access to data about us is systematically blocked. The network builds a world where data are uncontrolled; we insist on technologies to restore that control. Like the content industry, we propose heavy, expensive technologies to clog the flow of data on the Internet's pipes. Our instinctual reaction is to find a way to block the flow of data, since controlling the flow of data (like controlling the flow of "copies") coheres with our instincts about privacy.

But if the content industry must begin to think about copyright without control over copies, then we privacy fanatics should begin to think about privacy without control over data. We should begin to ask, how can we achieve the values that privacy serves without systematically staunching the flow of data?

This is not an easy question to ask—openly at least. There is a raging pro-privacy norm in the circles I frequent. The punishments for deviation from the party line can be quite severe. There is a set of fundamental truths that echo and are reinforced—that national IDs are bad, that government access to private transaction data is totalitarian, that medical records are sacrosanct—and the cost of questioning any of these truths is banishment from the pro-privacy club. Put differently: the only people who question these basic truths are people who don't really care about privacy.

I believe that this unwillingness to question is a mistake. And I believe that just as there is a cogent—indeed true—view of copyright that insists we can secure the values of copyright without controlling copies, there is a cogent—I'm not yet sure whether it is true—view of privacy that insists we can secure the values of privacy without controlling data. Indeed, the form of the right solution in each is similar, just as the form of the wrong solution (controlling data) in each is similar. Lessons from one can therefore suggest lessons for the other.

* * *

To make the argument work, however, we should begin by motivating the exercise. Why should we even look for a solution to the problem of privacy that differs from the solutions of the past? Why should we struggle to secure the values of privacy in a way different from controlling data?

The reason is similar to the argument advanced in the context of copyright: We lose, I would argue—in a different paper at a different time—an extraordinary potential when we build technologies to invert the basic technology of the Internet—which is to facilitate copying. The spread of knowledge through a free and unencumbered Internet, the advancement of culture through the free and unencumbered ability to remix culture, the potential for democracy in the powerful speech that these technologies can enable: these are the values that copyright through control gives up. And so if we could find a way to achieve what copyright seeks without losing these values, there's a reason at least to try.

The same is true about data. For my sense is that in our romance for privacy, we systematically undercut the extraordinary potential that free access to data might create. It is not just—or most importantly—the gains that are possible in the war against terror. It is instead the astonishing gains we could realize in a host of social contexts—from health care to economics—if we could cheaply gather and process the massive amount of data that digital technologies could produce. It is only through access to a massive amount of data about, say, DNA that we could begin to map links between DNA and disease. It is only through access to a massive amount of data about the health of every American that we could begin to identify and stop dangerous influenza before it becomes deadly.

The value of such access is obvious when questions of privacy are not drawn into the mix. Two examples suggest the point.

In a story published last April, a Wisconsin woman confessed to having staged her own abduction, as a way to get attention from her boyfriend.[3] Police found her bound and gagged in a swamp, her mouth taped with duct tape. But when they also found store videos recording her purchase of duct tape and rope the day that she went missing, they were a little suspicious. They relied in their search upon an increasingly familiar part of modern life—devices that record data endlessly—yet the story in the *Times* recounting the investigation didn't even raise the question of privacy that recording purchases at checkout counters might present.

---

[3] Jim Adams and Dick Meryhew, "Student Charged in Staging Abduction; A Troubled Romance May Have Spurred Audrey Seiler's Alleged Deception," *Star Tribune* (Minneapolis, Minn.), 15 April 2004, at A1.

Or again, a very popular television series—*24*—is recounting in real time (each episode is an hour; the series has twenty-four episodes) a day in the life of a counter-terrorist unit, faced with a biological terrorist threat. This "CTU" has apparently unlimited access to every bit of transaction data about everyone, which it uses to sift through the puzzle of who is behind this catastrophic threat. Yet the question of privacy doesn't even present itself in the context. I suspect even Jeff Rosen—were he to stoop to watch television—would feel assured by the efficiency of these data searches.

These particular examples are just the beginning of a story about the good that data will increasingly be able to do. Data attached to powerful computation will give us an understanding of both good and bad in this world that we don't currently have. We might argue about how much good it can do, but I don't think that anyone in good faith could argue that it would not do good. The essence of the argument in favor of privacy—where privacy means control over data—is that whatever good it does, total awareness would also do severe harm.

But of course, as a privacy advocate myself, I don't disagree that total awareness without more would do terrible harm. My argument is not that data are good, so limits on how one uses or accesses data are bad. Instead, the argument I am advancing here is that total awareness is good, and the bad from total awareness can be avoided, without destroying the possibility for total awareness itself by insisting upon tools to protect privacy that try to block access and use of data.

* * *

I'm not the first to suggest we find different ways to protect the value of privacy, given the reality of new technology. The most influential example of this argument is by David Brin, who argued in his extraordinarily important book, *The Transparent Society*, that there's no way to avoid these technologies of surveillance, but that so long as surveillance is available equally—so long as I can spy on the government just as the government spies on me—we don't need to worry about the consequences of these technologies.[4] Brin offers transparency as the solution to, well, transparency. He rests security upon the hope that compensating norms will evolve. A different modality of regulation—norms—would provide, Brin believes, the necessary protection for privacy.

Before Brin, it was Louis D. Brandeis who most famously suggested that the value of privacy could be protected through a new legal device,

---

[4] David Brin, *The Transparent Society: Will Technology Force Us to Choose between Privacy and Freedom?* (Reading, Mass.: Addison-Wesley, 1998).

given the emergence of technologies that rendered the old device useless. Indeed, the "Right to Privacy"[5] that Brandeis and Warren argue for is a direct response to the failure of the then-existing doctrinal device for protecting privacy—property. It was strong rules of trespass that the common law relied upon to keep people private. It was an absolute and automatic system of state copyright law as well. But as communities became more dense, and as access to property became more skewed, the device of property as a tool for protecting privacy had begun to falter. Brandeis and Warren thus offered a different legal device—a cause of action sounding in tort—to achieve the same social ends.

It is Brin and Brandeis's method that I mean to follow here, though I would not accept the particular innovations that each suggests. Privacy through transparency assumes a tolerance that is just not extant; it ignores dynamic pressures that would systematically weaken important aspects of social life. And neither would more tort law suffice to remedy the particular mix of troubles that digital networks will create. Instead, we would need something more, and different, if the values of privacy were to survive in a world that gave up restricting data.

* * *

So what would that something more be? We don't yet know enough about how the architecture of the Internet will develop. But we can begin to imagine layers of protection architected into the network that would seek the same values that controlling data seeks, yet would not impose the same costs.

In the space left to me here, I want to sketch two principles that will suggest a more general strategy. Both rely upon a distinction that I will describe before introducing the principles.

The distinction is between a nym and a person—or alternatively, a name, and the thing named. A name of course is not a person—though if true, it is linked to a person. A person's character is not determined by his name—a fact that I, Lester Lawrence Lessig III, rejoice over almost daily. But names in real space are conveniently tied to people. They are crafted to be usable as easy links. If instead of Lessig, my name were XASF12d3242a2es112e1qe, that indicator might well link back to me, accurately. But it would not be a usable name—usable at least by humans.

And that's the key to a nym. A nym is a name usable only by a machine. It is arbitrarily complex so that it cannot be used except within the system of a technology. It may be arbitrarily long and, hence, not

[5] Samuel Warren and Louis Brandeis, "The Right to Privacy," 4 *Harv. L. Rev.* 193 (1890).

easily reproduced. Or it may be encrypted and, hence, changed as the environment changes. However it is built, the point is that this nym can have a life without obviously or easily being tied back to me. If it is tied back to me, then we say I have been "traced" through my nym. If it is not traced back to me, then my privacy is protected even though data my nym carries are revealed for all to use.

So against the background of this principle, consider two principles:

The first is a principle of *regulated traceability.* The mere fact that all the data about everything everywhere exist is not intrinsically troubling. Some believe there's a god who is omniscient. That fact doesn't make life for the neighborhood snoop any easier (unless God talks to the snoop.) The danger instead comes from data's traceability—the ability to link some fact to some person. That's the potential cost of pervasive data: that the data get linked to a person.

Yet this cost is contingent. For the benefit from pervasive data can often be achieved without effecting traceability. We can learn what we need to learn about patterns of disease without knowing who in particular has what disease. Data can teach without revealing their source. Or more accurately, we could architect a system of pervasive data without also building a system that easily links data to a particular individual.

One such architecture would assure anonymity for the individual whose data were used. Anonymity here means that there's no way to trace the data back to the individual. Another architecture could preserve traceability, or the possibility of traceability, while still making traceability difficult.

It is this second alternative that I'd like to pursue. For sometimes, of course, the ability to link data back to an individual is a benefit. It was a benefit, for example, to society that the police were able to link data about purchases to a girlfriend who cried for help. And it sure feels like a benefit when the federal agents on *24* are able to link the telephone call of the terrorist back to a particular individual. The question is whether we can preserve the benefits of traceability while also assuring that it doesn't become so easy to trace that any privacy is lost.

The solution, as many have argued, is not to regulate access to data, but to regulate the traceability of data. This can be achieved in two very different ways. The more familiar is regulation—legally enforced rules that say who can and who cannot get access to these data under what circumstances. The more successful would be technology—software and hardware code that says who can and who cannot get access to this trace. Both techniques are possible, and no doubt the ideal solution would mix the two.

In this essay, my aim is not to describe that mix. It is instead to point out one important corollary that follows from this analysis: IDs,

even "national IDs," help regulate (as in limit) traceability. For a properly architected ID not only authenticates the link between a person and a nym, but also limits the conditions under which such a link is revealed. A poor ID is poor either because it fails reliably to link an individual to a nym, or because it fails reliably to control the conditions under which such a link is made. A good ID is good because it reliably links a person to a nym, and reliably restricts the conditions under which such a link is made.

For example: Imagine you are stopped by the police on a highway. You're asked for a driver's license, for the purpose of determining whether you are licensed to drive. Given the current architecture of driver's licenses, that certification is achieved with a token that is (relatively) hard to fake, and information on the token designed to link you to the token. Such information may include your height, weight, age, eye color, and address, and the police officer then matches that information to the person she sees. If it matches, then the presumption is that the token has validated your authority to drive.

But in this story, there's lots that the ID reveals beyond the authorization of the holder to drive. The home address, for example, is information used to tie the individual to the token, but an individual may not want to provide it generally. The same goes for the age of the driver. Or any other bit of data on the license. All of the information used to prove the individual is licensed to drive is information that may have other, unnecessary, uses that the individual may want to restrict.

So then compare the driver's license to a well-architected national ID. The police stop you. You put a black card into a reader the police officer has, and place your thumb on the thumb scan. The thumb scan authenticates you as the holder of the card. The card reveals to the police officer that the holder is authorized to drive. The appropriate information has been secured (authority to drive); no other information has been revealed (age, address, etc.). The better-architected ID thus assures traceability, but can also assure traceability does not degenerate to produce transparency.

A second principle is more traditional, and follows from the first. This is the regulation of use. Regulating traceability limits access to data that link data to an individual. Regulating use limits the use of those linked data. The essential threat to the values of privacy is realized when a nym is linked back to a person. It is therefore here that the core protection must be secured. Agencies both private and public that have secured access to the data that link a nym to a person must be heavily regulated to protect that link. No agency upon its own should be able to generate the link. No agency should be permitted to reuse the data about the link beyond clearly specified purposes.

Both principles, as applied to the problem of privacy, are the mirror image of the most familiar principles solving the problem of copyright. The essence of copyright protection in an environment of promiscuous copying is to maintain the link between a particular work and an owner. Securing that link between nym and person, and assuring that link is not later broken, is the key to assuring proper, if indirect, compensation. Privacy protection in an environment of promiscuous data is the same, though its valence is inverted. Securing—as in regulating—the link between nym and person is critical, and assuring that any link once drawn is not later misused is also critical.

One final observation before I stop. I don't believe that systems have a nature, as in a character that cannot be changed. But I do believe we should orient policy in light of basic characters that will not be changed, easily. The essence of the code of the digital infrastructure that increasingly defines our life is the ability to copy and gather data. The first is antithetical to copyright. The second is antithetical to privacy. But our choice should not be between the network and copyright/privacy. Our choice should be between the benefits of the network, and the values protected by copyright and privacy, implemented in a particular way. If we can realize those values without compromising the value of the network, we should. Or obviously we should. Yet strangely, the obvious remains obscure.