KEEPING THE INTERNET
OPEN • INNOVATIVE • FREE

www.cdt.org

CENTER FOR DEMOCRACY
& TECHNOLOGY

1634 I Street, NW
Suite 1100
Washington, DC 20006

DATE	10/17/14
TO	National Science Foundation
FROM	Center for Democracy & Technology
# PAGES	5

October 17, 2014
Attn: National privacy research strategy
NCO, Suite II-405
4201 Wilson Blvd.
Arlington, VA 22230

RE: Comments to the National Science Foundation on the "National privacy research strategy," Document Number 2014-22239.

The Center for Democracy & Technology (CDT) is pleased to submit comments in response to the National Science Foundation's (NSF) Request for Information (RFI) on developing and supporting a national privacy research strategy. We applaud the NSF for bringing attention to this critical issue and for supporting a comprehensive research strategy. There is an abundance of worthy and useful research endeavors that we would support, and these comments are not meant to be an exhaustive list but rather an illustrative sample.

As an advocacy organization, our work is driven and supported by the results of research on technology, the Internet, and the influence of both on society. CDT's mission is to preserve the user-controlled and edge-driven nature of the Internet and champion privacy, security, freedom of expression, and fairness. We support laws, corporate policies, and technology tools that protect the privacy of Internet users, and advocate for stronger legal controls on government surveillance. Advocacy of this nature is dynamic and broad. While our experts diligently review the research efforts on these topics, the research that is available is sometimes too narrow or not substantiated by a peer review process. As the RFI noted, technology often outpaces the storied institutions that advocates and academics rely on to provide fundamental answers to questions. A national privacy research strategy could help coalesce ongoing research efforts and ensure that redundant projects, where they exist, are in the spirit of confirmation.

MEMO

Impressive and substantial research institutions focused on important questions related to privacy are flourishing across the country. Most recently, danah boyd launched The Data & Society Research Institute in New York City to focus on "social, cultural, and ethical issues arising from data-centric technological development."[1] Some institutions have focused on technology policy for years, including Princeton's Center for Information Technology Policy[2] which has been hosting scholars who are conducting important cross-disciplinary work for nearly 10 years. Centers like these provide opportunities for inter-disciplinary discussion and research by bringing together sociologists, engineers, and lawyers to analyze technology with the level of expertise it demands. Supporting this kind of sustained convening and research nexus is a necessary pillar of a national privacy research strategy.

CDT believes that policymakers should make informed decisions based on substantial research and be guided by deliberate thought on issues concerning technology and privacy. We reject the idea that our society must make a choice between privacy innovation and national security, or that the loss of control over our personal data is inevitable. We support the NSF's effort to bring more attention and resources to generating thoughtful work on these issues.

1. Privacy objectives:

Among the wealth of important research topics in this landscape, there are a few CDT would like to emphasize as critical to the national privacy research strategy: privacy-preserving data mining, making Domain Name System (DNS) queries confidential, and supporting end-to-end privacy solutions.

In response to growing concerns over privacy and data mining, CDT recommends that the national privacy research strategy include support for development of multiparty privacy protocols. There are many sectors where different institutions hold data relevant to a research problem but cannot share or pool the data because of concerns for privacy of the individuals involved or because the data is proprietary. For example, hospitals may wish to share or combine patient data in order to jointly mine the information for medical research. Secure multiparty computation allows researchers to compute results of algorithms without pooling the data such that they only observe the final results of the data mining computation. Researchers at Carnegie Mellon described the importance of this work in a 2009 paper:

> "This question of privacy-preserving data mining is actually a special case of a long-studied problem in cryptography called secure multiparty

[1] Data & Society: http://www.datasociety.net/

[2] Center for Information Technology Policy: https://citp.princeton.edu/about/

> computation. This problem deals with a setting where a set of parties with private inputs wishes to jointly compute some function of their inputs. Loosely speaking, this joint computation should have the property that the parties learn the correct output and nothing else, even if some of the parties maliciously collude to obtain more information. Clearly, a protocol that provides this guarantee can be used to solve privacy-preserving data mining problems of the type discussed above."[3]

Gains of efficiency on secure multiparty computation would have a meaningful impact not only for commercial purposes but also for researchers who would have access to massively larger data sets with substantially fewer privacy concerns. This type of cryptography is time consuming and costly, but has the potential to realize huge gains for both industry and individual privacy. In particular, efficiency in this technology could provide answers to the intelligence community's desire to combine all of their data—something privacy advocates would not support without mathematical assurances that one person would not have too much power over those in the database.

Additionally, finding practical ways to standardize a system of making Domain Name System (DNS) queries confidential could realize some fundamental privacy gains. For example, Domain Name System Security Extensions (DNSSEC) and DNS-based Authentication of Named Entities (DANE) are two standards-based proposals to authenticate DNS queries, but they are still non-confidential. That is, while they protect against spoofing DNS responses, those responses are still sent in an unencrypted manner. Further, DNS leakage is a serious, fundamental limit to privacy – communications over the Internet may be carefully routed in one manner, but DNS queries may be routed differently and expose the end-users IP address. Making DNS queries confidential and private will likely require adding a Tor-like mixing process (where traffic is bounced around independent servers) on a much broader scale. This would have to be coupled with an encrypted query protocol that ideally would be based on private information retrieval, a sub-set of the more general multiparty computation problem mentioned above.

There is already an emerging trend of developing and supporting end-to-end privacy solutions (where the keying material is held only by the client) and usable privacy tools. These trends suggest that users are eager to find ways to control the collection and use of data about them and that companies are eager to design products to respond to that market. We support efforts to increase transparency and privacy by continuing to develop tools with privacy-enabling defaults and using elements of privacy engineering to ensure that privacy is deeply embedded in the design, rather than reverse-engineered later.

[3] Lindell, Yehuda, and Benny Pinkas. "Secure Multiparty Computation for Privacy-Preserving Data Mining." *The Journal of Privacy and Confidentiality* 1, no. 1 (2009): 59–98. Accessed October 1, 2014. http://repository.cmu.edu/cgi/viewcontent.cgi?article=1004&context=jpc.

2. Assessment capabilities:
Individuals are tracked so pervasively these days that concerns about privacy are starting to blur rather than intensify. The burden of proof has been on advocates to show that efforts to monetize every minute individual behavior have left us living in a constant state of observation. CDT believes that the conversation around privacy has been overly concerned with trying to prove harm to users and that there are substantial concerns simply with the collection of this magnitude of data.

Earlier this year, Justin Brookman and Gautam Hans wrote a paper titled, "Why Collection Matters: Surveillance As A De Facto Privacy Harm."[4] In it, they argue that consumers have a legitimate privacy interest in the collection of data, not only in its use or adverse effects as a result of it. They outline concerns including secondary uses, data breach, internal misuse, and chilling effects as evidence of the negative effect of data collection practices. This is an area where additional focus would help shift the burden of proof away from advocates and toward those collecting data to address concerns about harm. There are exciting, emerging research results[5] in privacy-preserving analytics that we believe could be bolstered and further incented by strategic research investment.

Additionally, there are some fundamental legal definitions and questions on which scholarship would be very productive, if not the most glamorous. For example, the legal question of who owns data that is collected about individuals is a fundamental point to many discussions of privacy. Additionally, there is no legislative definition of "big data" or many of the other terms that we use to describe the landscape we would like policymakers to regulate. Unifying these kinds of basic terms is a non-trivial research objective as it would help researchers, advocates, companies, and engineers use a common language to describe their policy position and innovation objectives.

3. Multi-disciplinary approach:
The pervasive nature of technology lends itself to research from all perspectives. The higher education system does not generally lend itself well to interdisciplinary studies, excepting students who can achieve all requirements of two different arenas. Due to the unique nature of researching privacy, the NSF should promote strong investment in the existing institutions focused on data, technology, and privacy as well as encourage more academic institutions to establish their own in-house research centers to encourage cross-discipline dialogue and research.

[4] Brookman, Justin, and Gautam Hans. "Why Collection Matters Surveillance As A De Facto Privacy Harm." 2014. Accessed October 1, 2014. http://www.futureofprivacy.org/wp-content/uploads/Brookman-Why-Collection-Matters.pdf.

[5] Hall, Joseph Lorenzo. "Having Your (Big Data) Cake and Eating It Too." October 17, 2014. Accessed October 17, 2014. https://cdt.org/blog/having-your-big-data-cake-and-eating-it-too/

This will signal that there is a career path in this type of research for students from multiple disciplines and increase the number of high-quality researchers focused primarily on privacy. Additionally, the NSF could host conferences that allow scholars, innovators, and advocates to convene and collaborate.

4. Privacy architectures:

Digital information that fits the description of big data tests our traditional beliefs around the responsibilities and challenges of analyzing and managing information. Several scholars have contemplated the ways in which datasets that have these features should be managed and regulated differently (given the unique power and risks of large datasets) and CDT embraces finding a new approach. The Federal Trade Commission (FTC) recently held a workshop on “Big Data: A Tool for Inclusion or Exclusion,” which brought together scholars and professionals from a number of fields to discuss discrimination and other adverse effects of using big data. There are useful frameworks in existing regulation that can serve as a guide for accountability and transparency regarding data, but they may not be sufficient to capture the degree of information collected and mined in the future.

In particular, the concept of transparency raises difficult questions in the big data and machine-learning context. The black box of algorithms is in some ways a necessary evil that promotes business growth and innovation, but it is not conducive to traditional interpretations of transparency. The way that transparency was conceptualized in the Fair Credit Reporting Act, for example, was to allow consumers to see the information that is held about them and to provide a way to correct it. This simply does not apply to a machine learning system. We need to help researchers redefine transparency and operationalize it in a way that gives users meaningful control over their data in order to ensure that all information about individuals, sensitive or not, is treated with respect. Several notable scholars are already doing this work, but additional support would encourage faster progress and perhaps lead to an operational result in the near future.

Thank you again for the opportunity to provide feedback to your questions. If you have any follow-up questions, please feel free to contact us at 202.637.9800.

Joe Hall
Chief Technologist

Alethea Lange
Policy Analyst