Meeting Between Council for Big Data, Ethics, & Society[1] and National Science Foundation

March 6, 2015
NSF Offices in Alexandria, VA

Attendees:

- Chaitan Baru (NSF)
- Kate Crawford (D&S)
- Alejandro Saurez (AAAS, NSF)
- Heng Xu (NSF)
- danah boyd (D&S)
- Amy Apon (NSF, computer systems research & BIGDATA)
- Martin Wiener (AAAS, NSF, Brain program)
- Jeryl Mumpower (NSF)
- Peter Arzburger (NSF, Cyberinfrastructure)
- Melissa Cragin (NSF, BIO/OAD)
- Geof Bowker (UCI)
- Steve Slota (UCI)
- Bonnie Tijerina (D&S)
- Renata Rawlings-Gross (AAAS, NSF BIGDATA)
- Sylvia Spengler (NSF, Program Officer)
- Jacob Metcalf (D&S)
- William Miller, Science Advisor, CISE/ACI
- Helen Nissenbaum, NYU

Top-Level Agenda:

1. Introduction to the Council
2. Existing and emergent ethical challenges that the NSF sees
3. Discussion: How can the NSF encourage ethical thinking in the proposal/funding process?
   A. One approach: Data Management Plans
   B. Other research based approaches; range of issues, disciplinary sources?
   C. Data Science Education
4. Future of the Council

Introduction & Summary:

[1] Funding for this Council was provided by the National Science Foundation (#IIS-1413864).

The purpose of this meeting was to put Council on Big Data, Ethics, & Society leadership in touch with a variety of NSF personnel to find opportunities for cross fertilization and new points of traction. danah boyd introduced the Council as a group of thought leaders from a wide range of disciplines trying to launch new conversations about ethics and big data practices. As the NSF increases investments in basic big data research in general, and the BIGDATA program in particular, it would be very useful to simultaneously find research gaps and advance ethics conversations. The Council has identified the recurring controversies around de-identified data and human subjects research protection as a core issue for funding agencies to grapple with. It has also identified the NSF's Data Management Plan requirements and new data science curricula as plausible points of leverage.

The NSF staff were largely receptive to the Council's plans and offered a number of possible routes for influencing policy and research agendas at the NSF and suggested additional avenues for traction, collaboration, and support. Much of the conversation focused on the need for infrastructure capable of fostering collaboration around emerging ethical challenges. All parties agreed that support for new data science curricula is critically important.

Meeting Outcomes:

The meeting began with a wide-ranging discussion of the state of ethics in the data sciences.

Early on, a few participants discussed how computer science researchers have historically had little contact with human subjects that required protection and thus computer science has operated largely outside of the formal ethics infrastructures and debates familiar in the social and behavioral sciences. However, big data techniques have put far more computer science researchers in contact with human subjects data and introduced new layers of ambiguity about what types of research require ethics oversight.

Several participants noted that there are important ethical questions for data science that lie outside of familiar human subjects protections. For example, environmental justice, money and credit are areas that are important to NSF projects engaged with Big Data that fall outside of the usual purview of human subjects protections. Peter Arzburger suggested that the Council focus on **collecting/producing concrete examples of data ethics problems in areas such as physics where practitioners are most unfamiliar with human subjects data.** All researchers have *some* contact with data ethics insofar as there are norms about how and how not to use data, but many do not have substantial experience considering data ethics in their own field beyond those generalities. It was noted that there is an important distinction between using data ethically and the ethical implications of big data techniques writ large. Jake Metcalf pointed out that **as data scientists move across traditional disciplinary boundaries they should be equipped to handle many kinds of data**—a physicist may very well come into contact with behavioral data at some point in their career.

One issue that Kate Crawford raised was whether data science education would uncritically enable problematic industry practices. She argued for developing curricular support that enables data science to come into contact with long-running similar debates in the humanities and social sciences. danah boyd said that corporate-academic collaborations have data ethics concerns, and **asked the NSF to share**

examples of how these questions—who to fund, how to provide ethical support—play out across NSF directorates.

Sylvia Spengler pushed attendees to think beyond human subjects; she offered examples of research that has to balance openness of data with the desire to prevent harm, such as data about threatened species that could disclose their locations. Martin Wiener echoed this by pointing out that people with PhDs in medical genetics cannot interface with patients because they lack the clinical training of MDs and therefore struggle with handling diagnostic data. Chaitan Baru raised the example of ecological science in Brazil, which requires all data to be stored on servers physically in Brazil due to mistrust of foreign researchers resulting from colonial exploitation. Baru also stated that there is **very little guidance available to data scientists struggling with issues outside of biomedical data**—if a researcher raises such concerns on campus there is rarely anyone to speak with. Several participants noted that data ethics face some significant challenges with **cross-border and cross-cultural ethics** because data infrastructures so often cross borders with different legal and ethical norms.

Helen Nissenbaum identified several threads in the issues raised so far. First, scientists and engineers might not realize that there is something ethical going on in their research.
Or, second, they may recognize there is an ethical issue that is going on but don't know where to resolve it. She argued that we need to also attend to a third case where science and technology change research practices such that our established ethical practices don't really hold up anymore. The Council could provide a way to help people work through the reasoning necessary to address such problems at the ground floor.

Conversation turned to the matter of how to **foster collaboration between ethicists and technologists**. Renata Rawlings-Gross noted that when technologists hear "ethics" they often think it means impediments to their work. Geof Bowker said that this results in researchers not wanting to reach out for help when they run into a problem, and that we need to **develop a good model for working *in situ*** with technical folks as they build out systems and infrastructures. Kate Crawford noted that there is more and more research showing that raising ethical questions early makes research and design better.

Crawford cited a growing schism between technologists and social scientists around the uses of human data and the appropriateness of consent. Sylvia Spengler shared a recent case under review at the NSF where the researchers explicitly wanted to use foreign social media data because they felt it would not require the same IRB review as domestic social media. However, the use of the foreign data would undermine reproducibility of the results. In effect, the **less rigorous ethical review made the science weaker**. Chaitan said this **balancing act between reproducibility and consent is a way to tie together research practices and ethics.**

danah boyd turned the conservation toward how the Council might help the NSF, raised the matter of DMP's, and asked for an update on the status of DMP revisions. Melissa Cragin discussed the efforts in the Biology Directorate to get an overview of how DMP's are constructed, and how they have changed since first being instituted in 2011. DMP's are not uniformly structured, posing a problem for retrospective analysis. They have found that in scientific communities that have long been handling large data sets, PI's are fairly articulate about the data management practices. However, research communities

that are newly engaged in big data techniques—and therefore do not have robust shared standards or infrastructure—have a harder time describing what happens to their data after the research project ends. The NSF sees a need to **establish better community standards for trust, security, and re-use across many sub-disciplines**.

Helen Nissenbaum asked how researchers are being asked to ethically vet their data. Chaitan Baru responded that DMP's should have more follow up to make sure they were followed. **DMP's also only address output of data—where/how it will be shared and stored—rather than input.** The NSF requires that by the time projects make it across their desks they have been cleared by local ethics oversight bodies, typically campus IRBs. Heng Xu said there is an increasing need for campus IRBs to be better equipped to deal with data ethics outside of biomedical data. She also expressed concern about whether corporations were adequately addressing data ethics in their research projects. Geof Bowker pointed out that people in the social sciences have long suffered under IRBs that are composed of untrained generalists rather than peers, and there is a wide variation in how data ethics is handled. danah boyd expressed enthusiasm for reforms of IRBs, but pointed out that **IRBs are dialed into very particular kinds of human subjects data and human subjects harms.** IRBs also do not offer the type of engagement that can help researchers thoughtfully restructure their projects.

Conversation turned toward the ways in which big data techniques put protected classes of data in contact with less-sensitive or unprotected classes of data. For instance, HIV researchers looking to correlate health status with social media content. There is little legal framework for handling such situations. Kate Crawford suggested that the ethical challenges of Big Data are mostly about new kinds of data or new relationships between data, rather than the scale of datasets suggested by the descriptor 'big'.

Melissa Cragin drew the conversation toward education. She noted that computer scientists are increasingly leaving academia for industry precisely so they can have more unfettered access to datasets. We therefore **shouldn't rely on university-based ethics mechanisms to enforce ethics, but instead build curricula that will reach data scientists early in their career**. Chaitan Baru seconded this, and shared that the new US Chief Data Scientist (DJ Patil) will be working with the NSF's Education Directorate to orchestrate new data science curricula. He described data science as a three-legged stool: computer science, statistics, and ethics/policy. Several people agreed that emerging data science curricula are important points of contact for the Council and the NSF. Helen Nissenbaum advocated for including an ethics component in capstone projects for data science degrees. danah boyd said that young data scientists need to grapple with controversial data sets, not just sanitized and vetted ones.

danah boyd asked that we consider a five-year timespan. How do we build a strong network of people that can think through these problems and cope with pitfalls of data science? The members of the Council will all do their own brilliant research, the question is how do we coordinate them around a coherent melody? Kate Crawford said the Council can produce a much more rigorous approach to data ethics and avoid top-down mandates. Geof Bowker said that he is less interested in best practices than best processes. Renata Rawlings-Gross suggested as a general strategy that we should **focus on ethically important practices that can be concretely addressed through research**. For example, the combination of disparate datasets is a common technical and ethical problem across disciplines.

The meeting concluded with many **practical next-steps:**

- Chaitan Baru said that the NSF would be hosting a mandatory BIGDATA PI workshop within the next year and encouraged the Council to participate in the agenda.
- Heng Xu suggested that some types of solicitations be revised to require collaboration with ethicists.
- Melissa Cragin said that the I-School movement is a good entry point, and a lot of I-Schools are just starting to spin up. Several people noted that the I-Conference often collects all of the I-School deans, providing an opportunity to pitch new projects.
- Martin Weiner said the Center for Science & Engineering Statistics will be ramped up quite a bit to provide more data-driven advice to Congress.
- We need social science research that can figure out how ethics is emerging in data research communities.
- Participate in the Council's/Data & Society's public conversations.
- Produce model curricula and other educational support.
- Build an engagement calendar to see each others' activities.
- Research Coordination Network (RCN) grants are designed to support an emerging discipline over a five-year span. Data ethics would be a ripe area for an RCN, and the Program Directors present strongly encouraged such an application.
- Need an Intra-Agency working group to coordinate data ethics.
- CISE has released an RFI about building regional hubs for big data research support.