RDA Panel Questions

What do you see as some of the major issues and concerns around data science/analytics that are NOT about privacy?
At a very high level, I'd say:

1. Lack of Transparency
    a. Do people who give their data know exactly what it will be used for? Is that clearly communicated?
    b. What other datasets will be combined with theirs to create a profile of a person?
    c. What other uses of their data will there be going forward?
2. Lack of Context
    a. We see many examples of data collected without understanding the populations whose dat is being collected or without considering underlying assumptions or prejudices that could be in the data. I've seen recent examples of data for social good organizations actually sending their data scientists into communities being studies and talking to real people before analyzing data. But, since data science is often seen as creating efficiencies, doing this level of due diligence is not happening.
3. Algorithms as Neutral and Outputs as The Truth
    a. Related to #2, a major concern I have is the lack of questioning the data - what bias might already being in the data? What bias might be baked into an algorithm since algorithms are made by humans who do have biases.
    b. I am also concerned about the assumption what gets spit out is the truth

All of these can impact fairness and potentially perpetuate discrimination of entire groups of people - in housing, jobs, and health services for example. Most industries are using data science to create efficiencies or improve services so it's critical that data science and analysis is thought about through an ethical lens or it will have even bigger impacts in the future.

In the academic environment, we did exploratory research that helped us uncover ethical issues that surface throughout the research project by interviewing computer science and data science researchers at US schools with top-tier CS programs. The researchers, faculty, and grad students include issues with data collection, data storage, and data reuse. For me, when it comes to data research in academia, I'd say major issues are:

1. Lack of proper oversight.
    a. Often when I talk about Ethics in data research, people say that's the responsibility of IRB and stop the conversation there. But, the researchers we interviewed say:
        i. IRB is not tech savvy enough. They may not know what they are looking for. A researcher sees them as a formality that needs to be checked off a

list, not as a resource to help as potential ethical issues arise throughout the research process.

ii. My colleague at Data & Society, Jake Metcalf explained how traditional ethics oversight may not work for this new type of research: “Research ethics regulations are so different in the data science and internet research world. For example, long-standing research ethics regulations typically exempt from further review research projects that utilize pre-existing and/or public datasets, such as most data science research. This was once a sound assumption because such research does not require additional intervention into a person’s life or body, and the ‘publicness’ of the data meant all informational or privacy harms had already occurred. However, because big data enables datasets to be widely networked, continually updated, infinitely repurposable and indefinitely stored, this assumption is no longer sound—big data allows potential harms to become networked, distributed and temporally stretched such that potential harms can take place far outside of the parameters of the research.”

2. How ethics is taught is still a work on progress
    a. Seperate classes are not always the best method
    b. Embedding in technical courses includes regular reminders of concerns
    c. Ethics is not just a checklist, it’s a way of thinking, questioning, and seeing the world. It’s part of a conversation with peers. That needs to be taught and it’s not that easy.

What are steps that we can take to develop a culture of ethics in corporate settings, where so much of the data analytics work takes place?

1. Making a culture of ethics means it is prioritized throughout the organization.

Make it not just a top-down issue but an organization-wide issue that requires involvement at all levels, including employees, middle management, and top execs.

Incorporating into mission statements and work processes.

2. Process: Have a real process, checklists, teams or committees who are responsible for answering questions or creating a rubric around when, how, and why data is collected, then examining assumptions. Several frameworks exist and can be used to help with these kinds of conversations.

How do we develop a code of ethics/checklist for ethical uses of data science? Should we? And who is “we”?

First, I think there are plenty of codes of ethics. The more practical stuff is what needs to be available now. Many organizations and individuals have already have created toolkits, guides, and checklists for specific communities. It’s best to look at the those that do exist and see where another would be useful. Also, are you talking discipline-specific or just a general one? I know needs and techniques are different for different academic disciplines.

Some examples of checklists include:

Deon - An ethics checklist for data scientists created by drivendata.org
Deon is a command line tool that allows you to easily add an ethics checklist to your data science projects. This checklist is designed to provoke conversations around issues where data scientists have particular responsibility and perspective. This conversation should be part of a larger organizational commitment to doing what is right.

For academic research using data science:
Networked Systems Ethics Guidelines created by Bendert Zevenbergen, Oxford Institute

These guidelines aim to underpin a meaningful cross-disciplinary conversation between gatekeepers of ethics standards and researchers about the ethical and social impact of technical Internet research projects. The methodology guides stakeholders to identify and minimize risks and other burdens, which must be mitigated to the largest extent possible by adjusting the design of the project before data collection takes place. The aim is to improve the ethical considerations of individual projects, but also to streamline the proceedings of ethical discussions in Internet research generally.

The primary audience for these guidelines are technical researchers (e.g. computer science, network engineering, as well as social science) and gatekeepers of ethics standards at institutions, academic journals, conferences, and funding agencies.

As an example, here are questions that are asked. These questions may also serve as a starting point for ethics committees of a department, journal, or conference, for their internal considerations:

- Context: How would you describe the context within which data is collected, information flows are created (or affected), or phenomena are measured?
- Aims: What are the aim and benefits of the project?
- Benefits: Why are the benefits good for stakeholders?
- Purpose limitation: Can the scope of data collection be limited whilst still achieving the project aim?
- Politics and Power: Are particular stakeholders empowered or disempowered as a result of this project?
- Risk of Harm: Could the collection of the data in this study be reasonably expected to cause tangible harm to any person's well-being?
- Law: Which bodies of law are likely to be applicable to the operation of the project?
- Values: Which values will the project conceivable impact?
- Burdens: Who carries the burden of harms or impacted values, and how?
- Technology Ethics: Can the harms and impacted values be traced to parts of the technological design of the project?
- Function Creep: Does the project potentially set a precedent for unethical methodologies that could be misused by others in the future?
- Data Governance: Using current techniques, can the data used in this study reveal private or confidential information about individuals?
    - If so, discuss measures taken to keep the data protected from inappropriate disclosure or misuse.

- Data Retention: When will the collected data be deleted?
- Tech Alternatives: Have you considered measures to mitigate the identified risk of harm or impacted values?
    - Can alternative technologies be employed or devised to mitigate some issues?
- Limit Scope Can you limit the scope of the project (geography, knowledge generated, etc.)?
- Methodology: Have others used alternative methodologies to achieve similar ends?
- Informed Consent? Do you need to rely on informed consent from participants and stakeholders?

For civil society, University of Chicago's Data Science for Social Good created the Ethical Checklist for Data Science Projects.
https://dssg.uchicago.edu/2015/09/18/an-ethical-checklist-for-data-science/

What are mechanisms we can develop/are being developed to evaluate the outcomes of algorithmic decision-making?
Toolkits, Guidelines, Frameworks that support a discipline or support DS work.
Examples:
Ethics & Algorithm Toolkit - https://ethicstoolkit.ai/ - John Hopkins' GovEx, the City and County of San Francisco, Harvard DataSmart, and Data Community DC have collaborated on a practical toolkit for cities to use to help them understand the implications of using an algorithm, clearly articulate the potential risks, and identify ways to mitigate them.

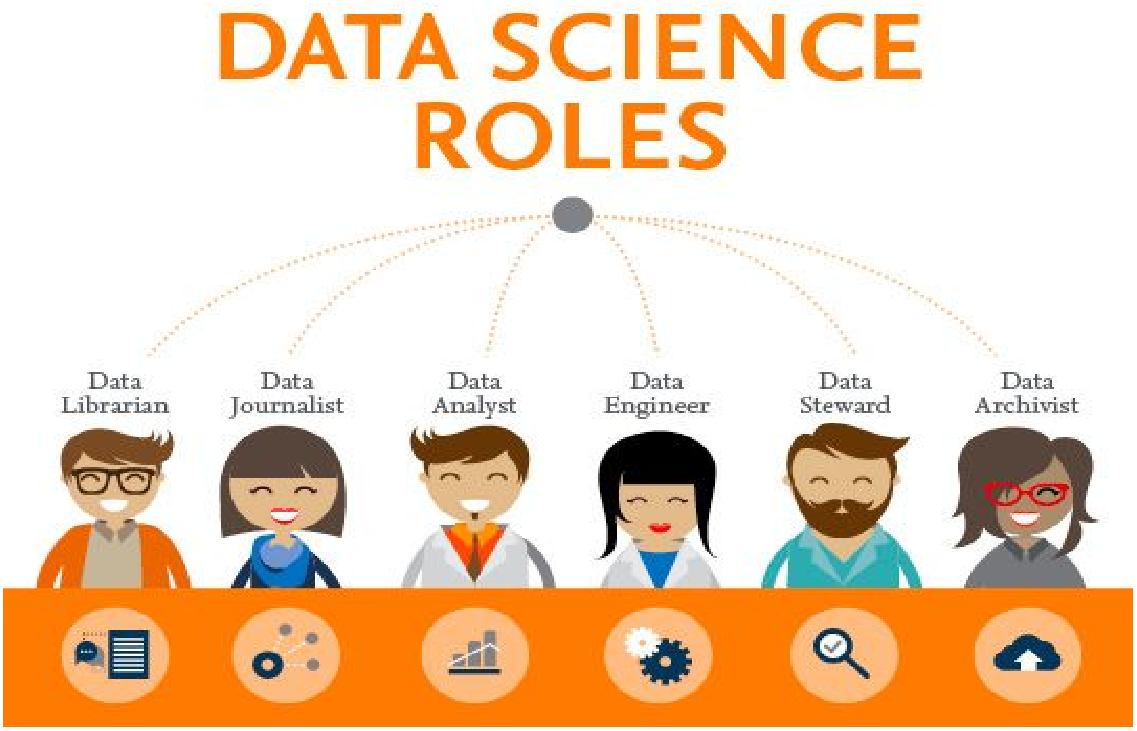
Ethical OS - https://ethicalos.org/ created by the Institute for the Future and Omidyar Foundation Creates risk mitigation questions to help people think about the tech they are building.

- **A checklist of 8 risk zones** to help you identify the emerging areas of risk and social harm most critical for your team to start considering now.
- **14 scenarios** to spark conversation and stretch your imagination about the long-term impacts of tech you're building today.
- **7 future-proofing strategies** to help you take ethical action today.

How does "the right to be forgotten" play into our concerns for data sources/outcomes of data analytics?
There needs to be ways to be taken out of data sets or opt out later. How can that be done? It impacts analysis as well as data storage and collection.

If we want to create diverse teams for data science, what should those teams look like?
It depends on the industry, I think. You may want a team with data scientists as well as someone tech savvy but with a deeper understanding of ethical issues that could arise. Depending on the work, you may want someone who comes from the group or groups you are targeting or gathering information from. There may be context missing that can only be explained by those in that group.

Dr. Liz Lyon has this imagine of different data science roles that may be part an a research institution. Having specialists who can go deeper may be better than just a team of data scientists and engineers. She includes:
Data Engineer, Data Analyst, Data Steward, Data Archivist, Data Journalist, Data Librarian



One of the major criticisms of algorithmic decision-making is the lack of transparency and accountability in such systems (examples include housing, financial products, health insurance). What are some ways we might put accountability into the system?

1. Government regulation - though we see this is not happening so far in the US.

2. Self-regulation - We see Facebook starting to do this, mostly because they are finally be scrutinized more closely. FB example from 3/19/19:

"Facebook on Tuesday agreed to overhaul its lucrative targeted advertising system to settle accusations that landlords, lenders and employers use the platform to discriminate, a significant shift for a company that built a business empire on selling personal data.

The settlement compels Facebook to withhold a wide array of detailed demographic information — including gender, age and Zip codes, which are often used as indicators of race — from advertisers when they market housing, credit and job opportunities.

Civil rights advocates have warned for years that Facebook's ads violated anti-discrimination laws because advertisers were able to use the data to exclude African Americans, women, seniors, people with disabilities and others. The Justice Department allowed a lawsuit to proceed last year over Facebook's objections, arguing that the company can be held liable for ad-targeting tools that deprive people of housing offers. The settlements resolve lawsuits and other legal challenges filed in recent years by the National Fair Housing Alliance, the American Civil Liberties Union, the Communications Workers of America and others.

The news is likely to reverberate through the tech industry. Google, Twitter and Amazon all offer similar demographic targeting tools, and companies such as LinkedIn have brisk businesses in employment recruiting. "Presumably every platform now will abide by the same terms of this settlement or risk being sued," Martin said."

3. Teach ethics so technical employees come in with a mindset.