Do No Harm: An Ethical Data Life Cycle

Thursday, April 4, 2019

Author(s)

Vandana Janeja

Ethics generally works on the principles of do no harm. Although research protocols to protect human beings have been in place for a while now, the pervasiveness of multiple types of data and their use make it less clear where the impact on human beings is in the data life cycle. Thus, harm is not only direct based on exposing identifiable data for individuals, but also indirect resulting from the reuse of easily available data and combining multiple datasets[1].

In particular for data science there is a need to develop ethical critical thinking while analyzing the data. Throughout the entire lifecycle of the data in the knowledge discovery process there are many opportunities for ethical decision making that a data scientist can evaluate to do no harm.
Here the harms are not only through identification of Personally Identifiable Information (PII), but also other types of data and algorithms that may not necessarily fall under the direct purview of traditional Institutional Review Board protections. These other types of data may indirectly lead to identifying human behaviors that should fall under privacy concerns. In addition, ethics may not always equate to privacy and there could be other types of data, which may not directly be identifiable but still have to be carefully handled. There may be yet other types of data which may have impacts for the society as a whole. Let us consider some examples that exemplify these ethical issues.

- Consider data such as location data collected on a phone, driving data collected through on board diagnostics (OBD), atmospheric data collected through sensors in a community. Now let us consider the types of patterns one may discover if they have access to such data. For example, the location data on the phone may pinpoint the location of secure locations where an individual is carrying the phone; OBD driving data may provide insights into driving behavior; atmospheric data may indicate the levels of toxins in the air. As we think through the examples we can also think of the uses of the data in conjunction with other types of data available. For example, if the data of atmospheric toxins is combined with geographical distribution of demographics, does that identify vulnerable populations? Can the driver behavior discovered through OBD or the information on the neighborhoods that the driver passes through be used to deny insurance to the driver or can the premiums be increased? Even in these limited examples we can see beginnings of questions of bias.

- Let us say you search for the word “Professional Haircut” in the image search of Google and observe the output. Now let us say you search “Unprofessional Haircut” in the image search of Google and observe the output. The difference in the two outputs is stark [2]. Similar, stark outputs can be seen with other contrasting keywords as well. While it is not clear if these outputs represent algorithmic bias or pervasiveness of the data with societal biases, data science students need to think through such issues and ensure they do not creep into the algorithms and create algorithmic biases or perceptions of biases.

- Consider the recent case of Cambridge Analytica[3], where not only personal data from users but also their friends were collected without their direct consent. Up until recently Facebook even had a feature that would simply allow for data scraping of its users just based on a simple search of their phone number or email address.

These questions of who is included or excluded, or how analyses are used, means that collaboration across multiple perspectives is crucial to good data science. Therefore, communication about work practices, including about data interpretation and data storytelling, are crucial to good data science. Data science is generally presented in the data life cycle pioneered through the KDD life cycle [4]. Using a traditional data life cycle as a backdrop, ethics cannot be tacked onto one part of the data life cycle but across the data life cycle infused into the process of discovering patterns in the data.
Such a modified and annotated lifecycle is depicted in the figure. This process has to start at **data collection**, beginning with the diversity of populations and privacy of individuals protected. During **data integration** ethical considerations play a key role in deciding whether data should be reused or combined with other datasets. Here questions should be carefully evaluated, such as: Should missing values be filled with averages from a broader population? Should anomalies be excluded from the data or included for a more careful assessment of the rare nuggets in minority data distributions?

Bias should be carefully avoided during **data selection** for the task being evaluated. **Data preprocessing** should consider sampling strategies, selection bias, as well as geospatial and temporal context in selecting the data and forming the right groups. In the next stage, in which data is processed through **pattern discovery**, algorithmic threshold choice can impact what patterns are discovered or excluded. Not only are thresholds an important consideration, but so are aspects of the provenance of the algorithm and the data, the reproducibility of the results, and the many other decision points during this phase with far-reaching ethical impacts. Once the patterns are discovered, it is essential that **pattern evaluation** include checks for detecting implicit bias and tested against ground truths to establish not only the accuracy but also veracity of the data and the results. Here veracity refers to trustworthiness of the data and results. If the data is imprecise and not representative, the results also will reflect this imprecision and non representativeness.

In addition to the ethical concerns at each step of the cycle, a design choice at one stage in the cycle can impact the following stages and result in compounded ethical issues. For example, if data collection is not representative of the population, then even if the algorithm is robust to bias, the patterns detected will be biased. So in essence, the model is only as good as the data and the patterns are only as good as the model. Knowledge gained from this cycle can go back into the cycle in a feedback loop, informing the data collection phase again.

With this process in place it is also important to remember the following principles of ethical thinking in data science:

- Ethics cannot be equated to privacy. Ethical decision making includes privacy, social responsibility, decision making, and evaluation of impact in an ethical framework.

- Ethics in data science can also include releasing of data, not just hiding it. More importantly, it includes releasing data responsibly with appropriate checks and balances in place. In addition, there should be transparency of the methods and processes to provide insights into how the data was prepared.

- Ethical thinking in data science means considering all of us and representing all of us in data.

- Ethical thinking in data science considers every data scientist touching the data as a data steward. This includes data collectors, data users, and data re-users.

- Ethical context is also heavily influenced and interpreted through a lens of other types of context such as space, time, activity.

With the advances in data science and the pervasiveness of data, asking the right questions in the
data lifecycle has never been more urgent.


Acknowledgment

The author would like to thank Dr. Lucy Erickson for many insightful discussions on ethical perspectives in data science and providing several suggestions to get this blog post into its current form. The author would also like to thank Dr. Susan Sterett for many perceptive conversations and shared resources on ethics in data studies.

Source URL: https://www.aaaspolicyfellowships.org/blog/do-no-harm-ethical-data-life-cycle

List of links present in page

- https://www.theguardian.com/technology/2016/apr/08/does-google-unprofessional-hair-results-prove-algorithms-racist-