A Day in the Life of an Ethical Data Scientist

How do data scientists spend their work time on the average day or week?

Data science depends upon fundamentals – having a clean and cohesive set of data to work with. One must search for any deficiencies, and either try to repair them, manage their limitations, or exclude them altogether. Without ensuring that these foundations are in place, any further processes are highly suspect and untrustworthy.

Once data is as clean as possible, then comes the validation process. Is this data likely to be a fair and accurate representation of reality? Is it pertinent to the questions that will be asked of it?

Having satisfied these areas, we can then proceed to an analysis of the data. This may be done manually, extracting insights and inferences using a combination of common sense, applied statistical methods, and perhaps also algorithmic models or machine learning.

How does a data scientist keep ethics a priority during their day to day work?

A data scientist should understand the provenance of data that they are using. Was it collected in a fair way, with appropriate consent? Has it been sufficiently anonymized to protect privacy? Could data be reverse engineered or uncovered through cross-correlation? What risks to individuals might arise if this process fails?

A data scientist should also be mindful of the elements that may constitute bias in their datasets. Has care been taken with sampling? Are any pre-trained models potentially bringing their own biases into one's ecosystem?

Another important point is that It’s easy to lie using statistics. Data can be ‘tortured’ to tell us almost anything we want to hear from it, with some effort. Therefore, another important aspect of ethical data science is to stand against any such requests that attempt to massage data into a format that does not reflect what it would tell us at face value.

Before they even work with the data, how do data scientists identify the questions or the problems they are trying to solve? How can they make sure they consider ethics at this stage?

Before even working with data itself, data scientists should consider if the intended purpose of the data is ethical, or if it might potentially be misused in an unethical manner.

It’s important that data scientists consider if the data that they are collecting for one purpose might also be applied to another (dual-use or multipurpose data). For example, data from grocery purchases might be used to recommend products, but it could also be used to infer gender, age, ethnicity or lifestyle, areas of life that users aren’t necessarily expecting to be exposed by their activities. Such data might be repurposed to create a profile on someone, which could end up influencing a risk classification system such as a credit scoring mechanism.

Not everyone speaks data, so what kinds of things to data scientists need to do to clearly communicate their work to all stakeholders?

It’s very important to communicate as clearly and simply as possible. Transparency and , explainability are key concerns when it comes to ethical emerging technologies. Transparency means not hiding things, and explainability (or explicability) means translating things in a manner that people can easily follow and comprehend.

Such openness naturally may need to be balanced with other needs for privacy and security concerns also. Transparency typically cannot be absolute. However, if it’s enough to provide informed consent as well as reasonable assurance, that’s generally sufficient for most people, aside from auditors or investigators.

Where can data scientists go to stay current about ethics and data science?

There are many excellent bulletins on the Ethics of AI from the likes of MIT Technology Review, The Montreal AI Ethics Institute, and Azeem Azhar’s Exponential View. These can help to keep one up to date with the latest developments in this fast-moving space.