CHAPTER 8

Data Ethics

The **panopticon** is a design that originated with the English philosopher and social theorist Jeremy Bentham in the eighteenth century. The device would allow prisoners to be observed by a single security guard, without the prisoners knowing they were being watched. Today, the panopticon is used as a metaphor to highlight the threat to privacy and personal autonomy that comes with the collection, processing, and analysis of big data and shows the need to protect personal information in the face of increasing technological advancement. For example, multinational businesses face increasing scrutiny over how to store, process, and transfer private user data across geographic boundaries[1].

In this chapter, we will discuss data ethics (derived from the Greek word *ethos* meaning habit or custom); the principles that govern the way we use, consume, collect, share, and analyze data; and how as practitioners of data science can ensure the decision systems we build adhere to ethical standards.

Although this might seem like a diversion from the technical into the realm of applied philosophy, what separates the data scientist from traditional software engineers is that we work with data and that data can represent real people or it may be used to make decisions about entire groups of people, for example, loan applications or to deny someone a loan.

[1] Increasing uncertainty over how businesses transfer data across geographic boundaries is a current issue in data ethics.

D. Sorvisto, *MLOps Lifecycle Toolkit*, https://doi.org/10.1007/978-1-4842-9642-4_8

If we don't consider what type of data goes into those decisions when we train a model, we may be responsible for building systems that are unethical. In the age of big data, where organizations collect vast amounts of data including demographic data which may be particularly sensitive, how that data is collected, stored, and processed is increasingly being regulated by policies and laws like the GDPR data protection act in the European Union and similar legislation in countries.

You may be a technical wizard at statistics or data analysis or software development, but without a solid understanding of data ethics, you may be doing more harm than good with your work, and your work may end up being a net negative to society as a whole. It is my opinion that what separates a scientist from a nonscientist or an engineer from a nonengineer isn't just knowledge but ethics. In this chapter, we will give a definition of data ethics and clarify some of the guiding principles you can use to shape your technical decision-making into ethical decision-making.

While this is not a book on generative AI, due to recent events, generative AI is set to shape a lot of the regulation around data ethics in the coming years. We will also cover some of the ethical implications of generative AI in this chapter, so you can understand the implications to your own organization if you incorporate generative AI into your data science projects.

Finally, we'll provide some recommendations for you how you can implement safeguards in your data science project such as retention policies to mitigate some of the risks that come with working with PII and other types of sensitive data.

Data Ethics

Data ethics is a branch of applied philosophy concerned with the principles that distinguish "good" decisions from "bad" decisions in the context of data and personal information. Unlike morality, which may determine individual behavior, ethics applies more broadly to a professional set of standards that is community driven.

Some of the ethical questions that data ethicists are concerned with include the following:

- Who owns data, the person it describes or the organizations that collect it?
- Does the organization or person processing the data have informed consent (important in the healthcare industry)?
- Are reasonable efforts made to safeguard personal privacy when the data is collected and stored?
- Should data that has a significant impact to society as a whole be open sourced?
- Can we measure algorithmic bias in the models we use to make decisions?

All of these questions are important as data scientists since we have access to vast amounts of personal data, and if this data is not handled properly, harm can be done to large groups of individuals. For example, an application that is meant to reduce bias in human decision-making but then exposes personal information of the groups it's trying to help may end up doing more harm than good.

Avoiding algorithmic bias and discrimination, improving the transparency and accountability of the data collection and analysis process, and upholding professional ethical standards are critical to the long term success of data scientists and MLOps and ensuring your models are valuable and sustainable.

Model Sustainability

I want to define the concept of model sustainability which I think is valuable to keep in mind when considering the role data ethics plays in data science and technical decision-making. What does it mean for a model to be sustainable? To be sustainable, it needs to adapt to change but not just technical change, change in society as a whole.

The fact is some data is political in nature; the boundaries between data and the individuals or group's data represents can be fuzzy, and when we start adding feedback loops into our model and complex chains of decision-making, how our models impact others may be difficult to measure. The other problem is social change is something not often considered by technical decision-makers, and a lot of software engineering is creating methodologies that protect against technical change but not social change. As data scientists, we need to be cognizant of both and have methodologies for making our models robust to social change as well which may come in the form of regulatory requirements or internal policy.

So how do we decide whether our model is able to adapt to regulatory requirements and social change? In the next section, we'll define data ethics as it applies to data science and discuss many issues around privacy, handling personal information, and how we can factor ethical decisions when making technical trade-offs.

Data Ethics for Data Science

How can we improve our ethical decision-making? In the real world, you may encounter trade-offs; for example, you may have data available that could improve the accuracy of your model. You may even have a functional requirement to achieve a certain accuracy threshold with your model. However, it's not acceptable to increase accuracy at the cost of algorithmic bias in the model. It is simplistic to assume just because a variable is "important" from a prediction point of view that it should be included automatically. This is also part of the reason why feature selection shouldn't be fully automated.

There are many ways to monitor bias, and this should be a part of the continuous monitoring process at minimum. A plan should be made to reduce algorithmic bias either by finding substitutes for variables that are sensitive, removing them all together. How the data was collected is also important; if the data was inferred without the user's informed consent, then it may not be ethical.

It's also possible stakeholders may not understand the implications of algorithmic bias in a model or the ethical implications of using sensitive PII in a model. In this case, it's the responsibility of the data scientist to explain the problem just as they would any other technical blocker.

Since data ethics is a rapidly evolving field, there are laws and regulations such as GDPR that can provide guidelines for making ethical decisions. In the next section, we will cover some of the most common legislation from around the world that may have impact to your projects.

GDPR and Data Governance

The General Data Protection Regulation (GDPR) is Europe's latest framework for data protection and was written in 2016 but became enforceable on May 25, 2018, replacing the previous 1995 data protection directive. The GDPR document has 11 chapters around general provisions, data rights, duties of controllers and processors of data, and liabilities for data breaches. One of the biggest impacts of the GDPR is its improvements in the way organizations handle personal data, reducing organization's ability to store and collect personal data in some circumstances and making the entire data collection process more expensive.

Personal data is any data which identifies or could identify a person and includes genetic data, biometric data, data processed for the purposes of identifying a human being, health-related data sets, and any kind of data that could be discriminatory or used for discriminatory purposes.

As a data scientist, if you do business with clients located in the European Union, you may have to abide by the GDPR. How does this translate into technical decision-making? You will likely have to set up separate infrastructure for the storage of data using a data center that is physically located in a particular geographic region. You will also have to ensure that when data is processed and analyzed, it does not cross this boundary, for example, moving data between geographic zones may have regulatory implications.

Similar legislation has been passed in several other countries since the GDPR such as Canada's Digital Charter Implementation Act on November 17, 2020. Although GDPR is a general data protection regulation (hence the name GDPR), there are regulations that apply to specific industries especially in healthcare and finance.

HIPAA: HIPAA or the Health Insurance Portability and Accountability Act is a 1996 act of the US Congress and protects patient data and health information from being disclosed. Since HIPAA is an American law, it only applies to American companies and when working with American customers, but if your organization does business with US citizens, you need to be aware of this law. The equivalent legislation in Canada is PIPEDA (Personal Information Protection and Electronic Documents Act) and is much broader than HIPAA, covering personal information in addition to health and patient data.

Ethics in Data Science

There are some guiding principles data scientists can use to make more ethical decisions. These principles include the following:

- Identify sensitive features and columns in a database and apply appropriate levels of encryption to PII (personally identifiable information).
- Set up mechanisms to decrypt PII if necessary but ensure that appropriate security and access controls are in place such as row level security and that only the information necessary to a job is made available.
- Add continuous monitoring to identify bias in model output for demographic data using metrics such as demographic parity.
- Assess the data set to understand if sensitive information could be inferred from any of the attributes, and take measures to remove these attributes or put in place appropriate safeguards to ensure this information is not misused.
- Understand how the models you develop will be used by business decision-makers and whether your model introduces any kind of unfairness or bias into the decision-making process either through the way the data is collected and processed or in the output of the model itself.

While these are not an exhaustive list, it should serve as a starting point for further discussion with your team to set standards for ethical decision-making and to highlight the importance data ethics plays in our own work. In the next section, we will look at an area that poses some risk for data scientists: the rise of generative AI.

Generative AI's Impact on Data Ethics

In 2023, Databricks released an ai_generate_text function in public preview that returns text generated by a large language model (LLM) given a prompt. The function is only available with Databricks SQL and Severless but can be used, for example, when creating a SQL query against a feature store. A data scientist could use this function to add generative AI to their project, and this is only one early example of how generative AI is increasingly making its way into data science tools.

The risk of being incorrect when discussing an event that is currently unfolding is relatively high, but this chapter wouldn't be complete without discussing the impact generative AI is having on data ethics. One of the biggest challenges generative AI poses to data ethics is related to data ownership.

How generative AI will impact how we view data ownership is still speculative as of 2023, but observers are already starting to see the profound impact it is having. A lot of the debate is around whether a human input into a model still owns the output of that model after it is sufficiently transformed. This is an incredibly interesting question that is poised to disrupt a lot of the current thinking that exists around data ownership, and if you use generative AI in your data science project, you need to be aware of the implications. I would suggest for the time being at least label output generated by a generative AI so you can identify it in your code base if you need to remove it in the future. Setting up a tagging system to this would be a clean way to implement a strategy in your own organization.

Safeguards for Mitigating Risk

We could spend years studying data ethics, and we still would never cover every scenario you might encounter. A compromise is needed between theory and practice to allow the reality of working with sensitive data attributes and PII and planning for the worst-case scenario such as a data breach or misuse of this information. Here are some safeguards you can implement in your own data science projects to mitigate this risk.

- Implement data retention policy, for example, removing data that is older than 30 days.
- Only collect data that is necessary to the model at hand and don't store data that is not relevant to the model especially if it contains PII.
- Encrypt all features that are considered PII such as email addresses, account numbers, customer numbers, phone numbers, and financial information such as credit card numbers.
- Consider implementing row level security and using data masking for tables and views that contain PII.
- Rotate access keys regularly and ensure data is encrypted in transit and at rest using latest encryption standards.
- Check PyPi packages and third party software before using them in a project in case they contain malicious software.
- Work with the security team to create a plan to monitor and protect data assets and minimize the risk of data breach.
- Implement continuous bias monitoring for models that use demographic data to ensure that the output is fair.
- Consider tagging anything created with generative AI during development.

Data Governance for Data Scientists

Data governance in the context of data science refers to a set of policies, procedures, and standards that govern the collection, management, analysis, processing, sharing, and access to data within an organization. Data governance is vital to provide guarantees that data is used responsibly and ethically and that decisions that come about as a result of a data analysis whether it be an ad hoc analysis or the output of an automated system are reliable, accurate, and ethical and are well-aligned with the ethical goals of the organization.

Data quality management is a part of data governance that implements data quality checks to ensure data is reliable. This goes beyond basic data cleaning and preprocessing and may include business initiatives in master data management and total data quality to maximize the quality of data across the entire organization rather than within a specific department.

Data security is another component of data governance and ensures that it is protected from unauthorized access and that the organization is taking steps to mitigate the risk of data breach. Policies such as requiring de-identification, anonymization, and encryption of data systems both at rest and in transit may be enforced by the data governance and security teams depending on the organization's threat model. The role of a data scientist and MLOps practitioner is to ensure the policies are implemented in accordance with these policies and to provide recommendations on how to mitigate risk of data breaches. Unfortunately, many data science tools are not secure, and malicious software is all too common in PyPi packages. A common attack is changing the name of a PyPi package to a name used internally by a data science team and hosting the malicious software on a public PyPi server. Such attacks are only the tip of the iceberg because security is often an afterthought in analytics and not a priority, even in Enterprise analytics software that should come with an assumption of security.

Data stewardship is another area of data governance related to data ownership but is more concerned with defining roles within different data teams such as data analyst, data engineer, MLOPs, and data scientist. In a RBAC or role based access control security model, each role would have well-defined permissions and responsibilities that can be enforced to protect data assets.

Finally, an organization should have a document defining and describing its data lifecycle. We talked about the MLOps lifecycle, but data also has a lifecycle, as it's created, and it's transformed into other data, creating new data sources, and these data sources are used but ultimately at some point are either deleted or archived and stored long term (requiring special consideration in terms of security). This entire lifecycle should be a part of the data governance process within your team to minimize risk of data loss and data breach and guarantee the ethical use of data across the entire data lifecycle.

Privacy and Data Science

Privacy concerns arise in the collection, storage, and sharing of personal information and data sources containing PII as well as in the use of data for purposes such as surveillance, voice and facial recognition technology, and other use cases where data is applied to identify individuals or features of individuals.

The history of data privacy can be traced back to the early days of computing. In 1973, some of the first laws on privacy were created with the passing of the Fair Credit Reporting Act which regulated the use of credit reports by credit reporting agencies to ensure not only accuracy but also the privacy of customer data. The following year, the United States also passed the Privacy Act which required federal agencies to protect privacy and personal information. Similar laws were passed in Europe in the 1980s, and by the 1990s with the rise of the Internet, data privacy concerns became an even bigger part of the public conscience.

Data science teams should only collect data that is necessary to the model at hand or future models and should ensure they have consent from the users whose data they're collecting. Not being transparent about the data collection process or how long data is stored means a risk to the reputation of the organization.

How to Identify PII in Big Data

When we're working with big data sets, these can be big in terms of volume but also in terms of the number of features, so-called "wide" data. It's not uncommon to have hundreds or even thousands of features.

One way to identify PII is to write some code that can dynamically churn through all of the features and verify columns like "gender," "age," "birthdate," "zip code," and any kind of demographic features that doesn't uniquely identify a person. While primary keys may be an obvious type of PII if they can be used to identify a person (e.g., a customer account key), for other features whether or not they can identify a person may require some more thought.

You may have to do some math around this; for example, if you have a combination of age or birth date and zip code, you might be able to identify a person depending on how many people live in a certain zip code. You could actually go through the calculation by using the Birthday Problem that states in a random group of 23 people, the probability of 2 people in that group having the same birthday is 0.5 or 50%.

We could generalize this heuristic and ask for any subset of demographic features in a big data set: What is the probability of a pair or combination of those features uniquely identifying a person? If the probability is high, you may have a hidden ethical trade-off between using the feature and increasing accuracy of your model and dropping the feature from your data set. At what specific threshold is acceptable to you and your problem depends on the problem, how the model is used, and your strategy for handling the PII.

This illustrates two important points when identifying PII: It's the combination of features that might uniquely identify someone rather than any one feature on its own, so when you're working with big data sets in particular, this is something to consider. Additionally, we can mathematically quantify the risk of identifying a particular person in a data set in some circumstances and actually quantify the risk.

Using Only the Data You Need

We've talked about PII but also there's a common sense approach here: We should only be using the data we actually need for the model at hand. Given, there may be an auxiliary need to use demographic data for marketing purposes, and that may be the reason why you need to include it in your feature set, but as much as possible, you should try to trim the fat and reduce the amount of data you're using. This also helps with performance; you don't want to bring in ten columns that are not needed since that's going to be a waste of space and bandwidth.

One question you can ask to trim the fat is are the features correlated with the response variable? This is a relatively common sense approach and may not work in all situations, but identifying the variables in your model that have no correlation with your target variable(s), you can create a shortlist of variables that could be removed. In the next section, we'll take a step back and look at data ethics from the point of view of data governance and the big picture impact of our models on the environment and society.

ESG and Social Responsibility for Data Science

Social responsibility and ESG (environmental, social, and governance) are increasingly becoming a part of organization strategies and future goals. Since data scientists seek to unlock value in data, understanding ESG and the role social responsibility plays in their organizations' long term goals will become increasingly important to the role of data science and MLOps.

Social responsibility in data science means the use of data to make decisions that benefit society, promote social good, and prevent harm to individuals or groups of individuals. An example is that data can be used to identify patterns of bias in big data and inform decision-makers on how these patterns of bias can be reduced. In industries such as energy, ESG involves a more concrete tracking of carbon emissions and the impact of the business on climate change and the environment and is a new opportunity for data scientists to drive positive change by coming up with innovative ways to measure ESG impact and make ESG initiatives data-driven.

Data Ethics Maturity Framework

If you remember way back in Chapter 1, we defined the MLOps maturity model and discussed different phases of maturity and how we could evaluate the maturity of a data science project. We can develop a similar framework for ethics in data science based on many data governance maturity frameworks used across industries.[2] Take a look at Table 8-1.

[2] Al-Ruithe, M., & Benkhelifa, E. (2017). *Cloud data governance maturity model.* https://doi.org/10.1145/3018896.3036394

Table 8-1. *Data ethics maturity framework*

Dimension	Questions	Definition	Level 1	Level 2	Level 3
Privacy	Is personal data collected, stored, and used?	Personal Identifiable Information (PII) is defined as information related to an identifiable person(s) such as email address, credit card, and personal address	Physical locks or physically protected PII, for example, in a safe or locked cabinet. PII may exist on paper	Digitization and digital safeguard such as RBAC or ACL. File level encryption or database encryption applied to PII columns	Both technological and policy safeguards. Administrative policy and governance to protect privacy
Bias (nonstatistical)	Do models make different decisions for different demographic groups? Could these decisions translate into unfair treatment for different groups or is the model fair?	Bias refers to different model output for different demographic groups and not bias in a statistical sense	Bias identified in model but no concrete safeguards	Bias identified and has been analyzed and measured. Plan to reduce bias in models	Continuous monitoring and active bias reduction in models

(*continued*)

Table 8-1. (*continued*)

Dimension	Questions	Definition	Level 1	Level 2	Level 3
Transparency	Is the data science team transparent about how data was collected?	Data transparency refers to data being used fairly, lawfully, and for valid purposes	What data sources are used is clearly documented	Business is aware of all data sources used and who has access to it	Individuals and businesses know which data sources are being used and data integrity is protected

How might you use Table 8-1 on your own project? Although we could add more dimensions such as social responsibility, data accessibility, and data security, understanding the impact of your data and models on transparency, privacy, and bias is a good starting point for understanding the ethical considerations around your problem.

Why a framework? It may seem like overkill but can help you to reduce technical risks associated with unethical use of data by providing a measurable and pragmatic method for evaluating and monitoring the project across these various dimensions.

This framework is not theoretical, and the process should also start early before data is collected since collecting and storing PII may already violate laws and regulations without having to have processed it. While transparency and privacy may also be qualitative dimensions that can't be measured directly, bias (not bias in the statistical sense) meaning whether the model is fair or not actually can be measured quantitatively using metrics like demographic parity (a kind of conditional probability). How you measure bias is different for each type of problem. For example, for a multi-class classification problem, you might compute bias differently than for regression, but continuous monitoring and active bias reduction would be what differentiates a level 3 and level 2 solution in this maturity framework. In the next section, we'll look at responsible use of AI and how some of the ideas around AI might apply to data science as a whole.

Responsible Use of AI in Data Science

Data science is not artificial intelligence, but data scientists may use AI such as generative AI both as developers to make themselves more productive and to generate features for even entire data sets. For example, one application of generative AI is you can sample from a generative model to "query" it, and this might be as simple as feeding in a prompt but could actually involve complex statistical sampling methods with applications from recommendation to data augmentation.

While the applications of generative AI in data science are without bound, there are ethical challenges posed by generative AI, and these are multifaceted challenges at the intersection of society, technology, and philosophy. It is not even known at this time whether emergent properties such as consciousness itself could arise from certain types of AI, and this creates a moral quandary.

With increasing attention on the responsible use of AI, the ethics of artificial intelligence is becoming a mainstay in many data science discussions across all types of industries and organizations, even those not traditionally seen as technology companies.

Topics like bias in large language models whether or not language models or other types of AI can have emergent properties like consciousness and the existential threat posed by AI will continue to push our understanding of data ethics.

Staying on top of the rapid advancements in AI almost requires a superhuman AI itself to digest the vast amounts of information available, but there are some resources available.

Further Resources

Data ethics is a rapidly evolving field and multidisciplinary field at the intersection of technology, society, and philosophy, so it's important to stay current. Some ways you can stay up to date with data ethics include the following:

1. Subscribing to industry publications including journals, magazines, and blogs and following news, for example, setting an alert for GDPR or similar data regulation. This may help to stay up to date on current debates and emergent news.

2. Joining reputable professional organizations. Although there is no centralized body for data ethics, finding like-minded professionals can provide guidance and practical experience that you may not find elsewhere especially if there is controversy around a particular ethical question.

3. Taking courses and reading the history of data ethics can help to make more informed decisions when working with data and personal information.

Data scientists and MLOps professionals that understand data ethics will help to standardize this body of knowledge and keep the data ecosystem free from long term negative consequences of making unethical choices when working with data.

In the final lab for this book, we will look at how you can integrate practical bias reduction into your project to reduce the risk of unethical use of the models you create, and the lab will provide some starter code you can use in your own project.

Data Ethics Lab: Adding Bias Reduction to Titanic Disaster Dataset

If you've done any kind of machine learning, you're probably familiar with this Titanic data set, but it's always struck me how people go through the example without thinking about the types of features used in the example, so it's always felt incomplete. In the lab, you'll add the necessary code to compute demographic parity to decide if the model is fair or not using the Shap library.

Here is the recipe for this lab.

Step 1. You'll first need to install the Shap library from PyPi preferably in your virtual environment.

Step 2. Run the python file chapter_8_data_ethics_lab.py.

Step 3. Decide if the model is biased or not based on the demographic parity. Feel free to change the data to make the model unbiased.

Summary

In this chapter, we defined data ethics and discussed why ethics are important for professionals that work with data including the following:

- Ethics for Data Science Projects
- GDPR and Data Governance
- Ethics in Data Science
- Further Resources

We looked at some guiding principles for applying ethical decision-making in data science and some case studies and examples of specific regulation that governs the ethical standards within the data ecosystem today. While technology and ethics are extremely important, it is only half the picture, and both regulation regarding data ethics and the technology we in MLOps are spared by domain knowledge and the intricacies of each individual industry. In the next chapter, we will look at specific industries from energy to finance and healthcare.