


Awarding
Great British
Qualifications




Introduction to Data Science and Big Data
Topic 11: Lecture 1
Data Science Ethical and Privacy Issues

Data Science Ethical and Privacy Issues Topic 11 - 11.2

Unit Syllabus

- Data Science and Big Data Fundamentals
- Introduction to Data
- Understanding Data & Exploration
- Data Pre-Processing
- Data Processing
- Model Selection and Evaluation
- Data Visualization
- Business Intelligence and Tools
- **Data Science Ethical and Privacy Issues**
- Unit Summary




Data Science Ethical and Privacy Issues Topic 11 - 11.3

Scope and Coverage

This topic will cover:

- What is Data Science Ethics?
- Accountability and Governance
- Data Provenance and Aggregation




Data Science Ethical and Privacy Issues Topic 11 - 11.4

Learning Outcomes

By the end of this topic students will be able to:

- Understand the ethical challenges and concerns in data science.
- **Understand the ethical challenges and concerns in data science.**
- Explore practical solutions for implementing responsible data practices.
- Recognize the importance of fostering a culture of data ethics within an organization.



Review Quiz

- 1. What is the primary purpose of Business Intelligence (BI)?
 - a. To replace human decision-making with automated algorithms.
 - b. To provide real-time streaming of data.
 - c. To extract meaningful insights from data to support decision-making. **Correct Answer (c)**
 - d. To focus exclusively on data storage.



Review Quiz

- 2. What distinguishes analytics tools from reporting tools in the BI landscape?
 - a. Reporting tools focus on historical data, while analytics tools provide insights for future trends.
 - b. Analytics tools are exclusively used for data visualization, while reporting tools handle data analysis.
 - c. Reporting tools are limited to basic data presentation, while analytics tools delve into advanced statistical analysis. **Correct Answer**
 - d. There is no distinction between analytics and reporting tools.



Review Quiz

- 3. In the context of data-driven decision-making, what does the intersection of BI and Data Science aim to achieve?
 - a. Eliminate the need for decision-making.
 - b. Enhance the speed of data visualization.
 - c. Provide comprehensive insights for informed decision-making. **Correct Answer**
 - d. Isolate BI and Data Science from decision-makers.



Review Quiz

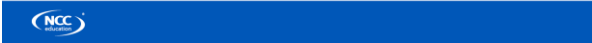
- 4. Which aspect emphasizes the synergy between BI and Data Science?
 - a. BI focuses on data storage, while Data Science focuses on data exploration.
 - b. The ability to seamlessly integrate predictive modeling into BI tools. **Correct Answer**
 - c. Data Science replaces the need for BI.
 - d. BI and Data Science are entirely separate and unrelated fields.



Data Science Ethical and Privacy Issues Topic 11 - 11.9

Last Topic

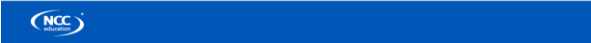
- BI analyzes historical and current data for structured decision-making, emphasizing reporting and data visualization.
- DSS offer modeling and scenario analysis capabilities for various decision types, including structured, semi-structured, and unstructured decisions.
- BI provides insights through data analysis and reporting, while DSS is more versatile and interactive, assisting users in the decision-making process with modeling and "what-if" analysis.
- BI tools are often used for monitoring key performance indicators (KPIs) and historical trends, while DSS systems are valuable for time-sensitive decisions and a broader range of decision scenarios.
- Organizations often use both BI and DSS tools to meet their diverse decision support needs, with BI informing structured decisions and DSS guiding a broader array of decision types.



Data Science Ethical and Privacy Issues Topic 11 - 11.10

What is Data Science Ethics?

- The moral principles, guidelines, and standards that govern the practice of data science.
- Involves making ethical decisions and considerations throughout the entire data science lifecycle,
 - includes data collection, data analysis, model development, deployment, and the use of data-driven insights.
- Essential because the use of data can have significant societal, economic, and individual impacts, and ethical considerations help ensure that these impacts are positive and fair.



Data Science Ethical and Privacy Issues Topic 11 - 11.11

Data Science Ethics

Ethical considerations in data science require a balance between extracting valuable insights from data and respecting the rights and dignity of individuals.

It's crucial for data scientists and organizations to incorporate ethical principles into their practices to ensure that data-driven decision-making benefits society as a whole.



Data Science Ethical and Privacy Issues Topic 11 - 11.12

Data Science Ethics Use Cases

OkCupid Data Scrape

- OkCupid is a U.S.-based, internationally operating online dating & friendship website .
- In 2016, almost 70,000 OkCupid profiles had their data released onto the Open Science Framework.
 - Two Danish researchers, Emil Kirkegaard and Julius Daugbjerg-Bjerregaard scraped the data with a bot profile on OkCupid
 - They released publicly identifiable information such as age, gender, sexual orientation, and personal responses to the survey questions the website asks when people sign up for a profile.



Data Science Ethical and Privacy Issues Topic 11 - 11.13

Data Science Ethics Use Cases

OkCupid Data Scrape

The two researchers didn't feel their actions were explicitly or ethically wrong, because "Data is already public."

What do you think about the ethics of releasing “already public” data.



Data Science Ethical and Privacy Issues Topic 11 - 11.14

Data Science Ethics Use Cases

OkCupid Data Scrape

The main concern raised was that even though data may be public, that doesn't mean someone consents to personally identifiable data being published on an online forum.

Ethically, not okay in the public's eyes.



Data Science Ethical and Privacy Issues Topic 11 - 11.15

Data Science Ethics Use Cases

AI “Beauty Contest”

- In 2016, the first AI (artificial intelligence) judged beauty contest selected 44 winners from the internet.
- The selection of winners raised concerns because of the 6,000 submitted photos from over 100 countries, only a handful were non-white.
- One person of colour was selected as a winner, and the rest of the non-white winners were Asian.
- The obvious problem with this, was that a majority of photo submissions came from Africa and India.



Data Science Ethical and Privacy Issues Topic 11 - 11.16

Data Science Ethics Use Cases

AI “Beauty Contest”

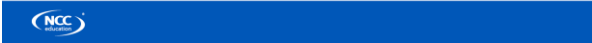
- The company, Beauty.AI said this project was a “deep learning” project that was sponsored in part by Microsoft.
- Beauty.AI claimed the algorithm used was biased because the data they trained it on was not diverse enough.
- For future projects, the hope is to correct the problem of bias by employing more sets of data and designing algorithms that can explain any bias.



Data Science Ethical and Privacy Issues Topic 11 - 11.17

Key Aspects of Data Ethics

- Privacy
- Fairness
- Transparency
- Data Ownership and Consent
- Beneficence
- Data Security
- Accountability
- Data Governance
- Legal and Regulatory Compliance

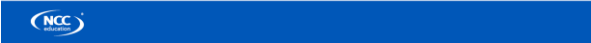


Data Science Ethical and Privacy Issues Topic 11 - 11.18

Privacy

The preservation of individuals' confidentiality, autonomy, and control over their personal information.

- Involves safeguarding sensitive data from unauthorized access or misuse,
- Obtaining informed consent for data collection,
- Ensuring data handling practices respect individuals' rights and expectations.



Data Science Ethical and Privacy Issues Topic 11 - 11.19

Privacy - Example

Healthcare Data Privacy:

- Healthcare organization collects and analyzes patient data for medical research purposes. Privacy involves :
- **Informed Consent:**
 - Before collecting any patient data, the healthcare organization must obtain informed consent from each patient.
 - Patients should be provided with clear information about what data will be collected, how it will be used, and who will have access to it.
 - Patients have the right to decide whether they want to participate.



Data Science Ethical and Privacy Issues Topic 11 - 11.20

Privacy - Example

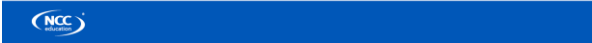
- **Data Anonymization:**
 - To protect patient privacy, the organization should anonymize or de-identify the data by removing or encrypting personally identifiable information (PII) such as names, addresses, and social security numbers. This reduces the risk of re-identification.
- **Data Security:**
 - The organization must implement robust data security measures to prevent data breaches or unauthorized access.
 - This includes encryption, access controls, and regular security audits.



Data Science Ethical and Privacy Issues Topic 11 - 11.21

Privacy - Example

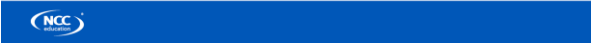
- Purpose Limitation:
 - Patient data should only be used for the specific research purposes for which consent was obtained.
 - Using the data for unrelated research or selling it to third parties without consent would violate patient privacy.
- Data Retention:
 - Clear policies should be established regarding how long patient data will be retained.
 - Once data is no longer needed for research, it should be securely deleted.



Data Science Ethical and Privacy Issues Topic 11 - 11.22

Privacy - Example

- Data Sharing:
 - If the organization intends to share patient data with external collaborators or researchers, it should do so under strict data sharing agreements that respect patient privacy and adhere to legal and ethical guidelines.
- Transparency:
 - Patients should be kept informed about the progress and results of the research, ensuring transparency in how their data is being used.



Data Science Ethical and Privacy Issues Topic 11 - 11.23

Privacy - Example

- In this example, privacy in data ethics ensures that patients' personal health information is handled with the utmost care, respecting their autonomy and protecting their sensitive data from unauthorized access or misuse.
- Privacy safeguards, such as informed consent, data anonymization, and strict data handling practices, are crucial to maintaining trust between the healthcare organization and the patients it serves.
- Failure to uphold these privacy principles could result in ethical violations and legal consequences.



Data Science Ethical and Privacy Issues Topic 11 - 11.24

Fairness

Preventing unfair or biased treatment of individuals or groups based on their characteristics, such as race, gender, age, or other protected attributes.

- Designing data-driven systems, algorithms, and decision-making processes in a way that minimizes bias and ensures equitable treatment for all.



Data Science Ethical and Privacy Issues Topic 11 - 11.25

Fairness - Example

- Credit Scoring Algorithm:**
 - a financial institution that uses a machine learning algorithm to determine creditworthiness for loan applicants. Fairness involves :
- Training Data Bias:**
 - The historical data used to train the credit scoring model may contain biases.
 - For instance, if the training data disproportionately approved loans for one racial group over another, it would introduce bias into the model.



Data Science Ethical and Privacy Issues Topic 11 - 11.26

Fairness - Example

- Algorithmic Fairness Metrics:**
 - The organization should define fairness metrics to assess the algorithm's performance in terms of fairness.
 - For instance, it could use demographic parity, which measures whether loan approval rates are roughly equal across different racial groups.
- Bias Mitigation:**
 - To address bias, the organization can employ bias mitigation techniques during model development.
 - This might involve re-sampling data to balance representations of different groups or using pre-processing techniques to debias the training data.



Data Science Ethical and Privacy Issues Topic 11 - 11.27

Fairness - Example

- Explanations and Transparency:**
 - Essential to make the credit scoring model transparent and explainable.
 - Applicants should understand why their application was approved or denied, and any factors used in the decision-making process.
- Ongoing Monitoring:**
 - The financial institution should continuously monitor the algorithm's performance to ensure it remains fair over time.
 - If new biases emerge, corrective actions should be taken.



Data Science Ethical and Privacy Issues Topic 11 - 11.28

Fairness - Example

- Fairness Audits:**
 - Periodic audits of the algorithm should be conducted by third-party experts to assess fairness objectively.
- Feedback Loops:**
 - Establishing feedback loops with affected communities and soliciting input can help identify and rectify fairness issues.



Data Science Ethical and Privacy Issues Topic 11 - 11.29

Fairness - Example

- In this example, fairness in data ethics ensures that the credit scoring algorithm does not discriminate against certain racial groups or other protected classes.
- By identifying and mitigating bias, providing transparency, and monitoring the model's performance, the financial institution aims to make credit decisions more equitable and just.
- Failure to address fairness concerns can result in systemic discrimination and legal challenges.
- Fairness is a crucial ethical consideration in data-driven decision-making, especially in areas with significant societal impact like lending, hiring, and criminal justice.



Data Science Ethical and Privacy Issues Topic 11 - 11.30

Transparency

Transparency in data ethics involves being open and clear about data practices, including data collection, processing, modeling, and decision-making.

- It ensures that individuals and organizations can understand and scrutinize the processes behind data-driven systems and algorithms.



Data Science Ethical and Privacy Issues Topic 11 - 11.31

Transparency - Example

Algorithmic Hiring System:

- A company that uses an algorithmic hiring system to screen job applicants. Transparency in this context involves several key elements:
- **Explanation of Data Usage:**
 - The company should provide clear and concise explanations to job applicants about what data is collected during the application process, how it is used, and how long it will be retained.



Data Science Ethical and Privacy Issues Topic 11 - 11.32

Transparency - Example

- **Algorithm Explanation:**
 - The company should make an effort to explain how the hiring algorithm works, including the factors and criteria used to evaluate applicants.
 - This could include the weight assigned to each criterion.
- **Fairness and Bias:**
 - Transparency also involves disclosing efforts to ensure fairness and mitigate bias in the algorithm.
 - The company should explain how it addresses potential bias in the hiring process, such as by using diverse and representative training data.



Data Science Ethical and Privacy Issues Topic 11 - 11.33

Transparency - Example

- **Decision Explanations:**
 - When an applicant is rejected or selected based on the algorithm, they should receive a clear explanation of why this decision was made.
 - This might include feedback on their qualifications relative to the job requirements.
- **Opt-Out and Data Deletion:**
 - Applicants should have the option to opt out of certain data collection practices and request the deletion of their data after the hiring process is complete.



Data Science Ethical and Privacy Issues Topic 11 - 11.34

Transparency - Example

- **Data Security and Retention:**
 - Transparency extends to data security and retention policies.
 - Applicants should know how their data is protected and for how long it will be stored.
- **External Auditing:**
 - The company might consider external audits or third-party assessments of its hiring algorithm to ensure transparency and fairness.



Data Science Ethical and Privacy Issues Topic 11 - 11.35

Transparency - Example

- By practicing transparency in this example, the company aims to provide job applicants with a clear understanding of the hiring process, reduce the risk of bias or discrimination, and build trust with potential employees.
- Transparency enables applicants to make informed decisions about whether to participate in the hiring process and ensures that they can access information about how their data is used and decisions are made.




Data Science Ethical and Privacy Issues Topic 11 - 11.36


Checkpoint Summary

- Data ethics involves the responsible and ethical handling of data, addressing privacy, fairness, transparency, and more.
- Solutions include data minimization, consent, transparency, bias mitigation, and accountability, while legal and regulatory compliance is crucial to protect individuals' rights.
- Embracing data ethics ensures responsible data use, maintains trust, and minimizes risks for individuals and organizations.





Awarding
Great British
Qualifications



Introduction to Data Science and Big Data
Topic 10: Lecture 2
Data Ethics

Data Science Ethical and Privacy Issues Topic 11 - 11.38

Data Ownership and Consent

Data ownership refers to the legal and ethical concept that individuals or entities have the right to control and make decisions about their own data.

It involves the idea that individuals are the ultimate controllers of their personal information.




Data Science Ethical and Privacy Issues Topic 11 - 11.39

Data Ownership and Consent

Social Media Platform: Suppose a social media platform wants to implement responsible data practices related to data ownership and consent:

- Data Ownership:** The platform acknowledges that users own their data, including their posts, photos, and personal information. It does not claim ownership over this data.
- Informed Consent:** When introducing new data collection or processing features (e.g., location tracking for geotagged posts), the platform seeks informed consent from users. It clearly explains the purpose of the feature, how the data will be used (e.g., to enhance content recommendations), and any potential implications (e.g., improved user experience). Users have the option to opt in or opt out of such features.




Data Science Ethical and Privacy Issues Topic 11 - 11.40

Data Ownership and Consent

- Privacy Controls:** The platform provides users with granular privacy controls, allowing them to adjust who can see their posts and profile information. Users have the ability to customize their privacy settings to align with their preferences.
- Data Deletion:** Users can request the permanent deletion of their accounts and associated data. The platform ensures that user data is deleted in accordance with privacy regulations.

By adhering to principles of data ownership and informed consent, the social media platform respects users' rights, autonomy, and privacy preferences. This ethical approach enhances user trust, helps prevent data misuse, and ensures responsible data handling practices.



Beneficence

- Beneficence in data ethics refers to the ethical principle of using data and technology for the benefit of individuals, society, and the greater good.
- It involves making data-driven decisions that promote positive outcomes while minimizing harm or potential risks.



Beneficence - Example

- Suppose a public health agency is collecting and analyzing data to address a public health crisis, such as a disease outbreak.
- Beneficence plays a crucial role in this context:
 - The agency collects data on infection rates, geographical hotspots, and demographic information, ensuring that data is comprehensive, accurate, and up-to-date.
 - Through data analysis, the agency identifies patterns, transmission routes, and risk factors associated with the outbreak. This information helps inform public health interventions and resource allocation.



Data Security

Data security is the practice of protecting data from unauthorized access, disclosure, alteration, or destruction. It involves implementing measures, policies, and technologies to safeguard data assets, ensuring confidentiality, integrity, and availability.



Data Security - Example

Healthcare Data Security: Consider a healthcare organization that stores electronic health records (EHRs) for millions of patients. Data security is paramount in this context:

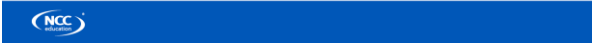
- **Data Encryption:**
 - The organization uses strong encryption techniques to protect EHRs both in transit (e.g., when transmitted between healthcare providers) and at rest (e.g., when stored in databases).
 - Encryption helps ensure that even if unauthorized access occurs, the data remains unreadable.



Data Science Ethical and Privacy Issues Topic 11 - 11.45

Data Security - Example

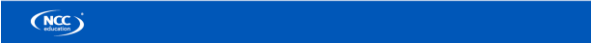
- **Access Controls:**
 - The organization implements strict access controls, limiting who can view, edit, or delete patient records. Access is granted based on the principle of least privilege, meaning that individuals only have access to the data necessary for their roles.
- **Authentication:**
 - Multi-factor authentication (MFA) is enforced to verify the identity of individuals accessing the EHR system, adding an extra layer of security beyond passwords.



Data Science Ethical and Privacy Issues Topic 11 - 11.46

Data Security - Example

- **Data Backups:**
 - Regular backups of patient data are performed to ensure data availability in case of system failures, natural disasters, or cyberattacks.
- **Security Audits:**
 - The organization conducts regular security audits and penetration testing to identify vulnerabilities and address them proactively.
- **Data Loss Prevention:**
 - DLP technologies are deployed to monitor and prevent unauthorized data transfers or leaks, both within the organization and externally.



Data Science Ethical and Privacy Issues Topic 11 - 11.47

Data Security - Example

- **Incident Response Plan:**
 - The organization has a well-defined incident response plan that outlines the steps to be taken in the event of a data breach. This includes notifying affected individuals and regulatory authorities as required by law.
- **Employee Training:**
 - Staff members are trained in data security best practices and are made aware of their role in safeguarding patient data.



Data Science Ethical and Privacy Issues Topic 11 - 11.48

Data Security - Example

- In this example, data security measures are in place to protect sensitive patient information from unauthorized access, data breaches, and other security threats.
- Ensuring data security is not only an ethical obligation but also a legal requirement in the healthcare sector, where maintaining the confidentiality and integrity of patient records is paramount to building trust and safeguarding patient well-being.



Data Science Ethical and Privacy Issues Topic 11 - 11.49

Accountability and Governance in Data Ethics

Accountability and governance are critical aspects of data ethics, particularly in the context of responsible data handling and decision-making. They help ensure that organizations and individuals are held responsible for their actions and decisions related to data.



Data Science Ethical and Privacy Issues Topic 11 - 11.50

Accountability

- Accountability in data ethics refers to the principle that individuals and organizations should be answerable for their actions and decisions related to data.
- This includes taking responsibility for the consequences of data use and ensuring that ethical standards are upheld.



Data Science Ethical and Privacy Issues Topic 11 - 11.51

Accountability

- Accountability involves several key responsibilities, including:
 - ✓ Taking responsibility for data collection, processing, and usage practices.
 - ✓ Acknowledging and rectifying any mistakes, biases, or harms that may arise from data-related decisions.
 - ✓ Being transparent about data practices and sharing information with relevant stakeholders.
 - ✓ Complying with legal and regulatory requirements related to data use.
 - ✓ Establishing mechanisms for redress and remediation when data misuse occurs.
- Accountability helps build trust with data subjects and ensures that ethical principles are not merely theoretical but are put into practice.



Data Science Ethical and Privacy Issues Topic 11 - 11.52

Governance

Data governance refers to the framework of policies, procedures, and practices that an organization puts in place to manage data effectively, securely, and ethically.

It encompasses the rules and processes for data collection, storage, access, and use within an organization.



Data Science Ethical and Privacy Issues Topic 11 - 11.53

Governance

Components of Governance

- **Data policies:** Clearly defined policies that outline how data should be handled, including privacy, security, and ethical considerations.
- **Data stewardship:** Assigning responsibility for data management to specific individuals or teams within the organization.
- **Data quality:** Ensuring that data is accurate, reliable, and consistent.
- **Data security:** Implementing measures to protect data from unauthorized access, breaches, or other security threats.
- **Compliance:** Ensuring that data practices align with relevant laws and regulations.
- **Ethical guidelines:** Integrating ethical principles into data governance practices.



Data Science Ethical and Privacy Issues Topic 11 - 11.54

Governance

- **Purpose:**
 - Data governance is essential to maintain the integrity and trustworthiness of an organization's data assets. It helps mitigate risks associated with data misuse, ensures compliance with legal requirements, and promotes ethical data handling.
- **Accountability** ensures that individuals and organizations take responsibility for their actions related to data
- **Governance** provides the structure and rules necessary to manage data in an ethical and responsible manner.

Together, they help create a foundation for ethical data practices and maintain trust with stakeholders, including data subjects and the broader public.



Data Science Ethical and Privacy Issues Topic 11 - 11.55

Legal and Regulatory Compliance

Legal and regulatory compliance in the context of data ethics refers to the obligation of individuals and organizations to adhere to laws, regulations, and standards that govern the collection, processing, and handling of data.

Compliance ensures that data practices align with legal requirements and ethical principles, protecting individuals' rights and minimizing risks.



Data Science Ethical and Privacy Issues Topic 11 - 11.56

Data Provenance

Data provenance refers to the documentation of the origin and history of data, tracking the sources and transformations that data has undergone throughout its lifecycle.

It provides a record of how data was collected, processed, and modified.



Data Science Ethical and Privacy Issues Topic 11 - 11.57

Data Provenance

Data provenance is crucial in data ethics for the following reasons:

- Accountability
- Transparency
- Trust

Data provenance is often employed in domains where data integrity, accuracy, and traceability are critical, such as scientific research, financial auditing, and healthcare, to ensure that data can be trusted and validated.



Data Science Ethical and Privacy Issues Topic 11 - 11.58

Data Provenance

Data provenance refers to the documentation of the origin and history of data, tracking the sources and transformations that data has undergone throughout its lifecycle.

It provides a record of how data was collected, processed, and modified.



Data Science Ethical and Privacy Issues Topic 11 - 11.59

Topic Summary

- **Data ethics** involves the responsible and ethical handling of data, addressing privacy, fairness, transparency, and more.
- Ethical data practices encompass **fairness, transparency, accountability, beneficence, and respecting data ownership and consent**.
- Ethical dilemmas in data science include privacy protection, fairness in algorithms, transparency in decision-making, and responsible AI use.
- Solutions include data minimization, consent, transparency, bias mitigation, and accountability, while legal and regulatory compliance is crucial to protect individuals' rights.
- Embracing data ethics ensures responsible data use, maintains trust, and minimizes risks for individuals and organizations.



Data Science Ethical and Privacy Issues Topic 11 - 11.60

Next Topic 12: Unit Summary

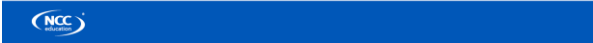
- Data Science and Big Data Fundamentals
- Introduction to Data
- Understanding Data & Exploration
- Data Pre-Processing
- Data Processing
- Model Selection and Evaluation
- Data Visualization
- Business Intelligence and Tools
- Data Science Ethical and Privacy Issues




Data Science Ethical and Privacy Issues Topic 11 - 11.61

References

- David, M. (2022). Data Science Ethics : Concepts, Techniques, and Cautionary Tales. Oxford University Press.
- Loukides, M., Mason, H., & Patil, D. (2018). Ethics and Data Science. O'Reilly Media, Inc.
- O'Keefe, K. & O'Brien, D. (2023). Data Ethics: Practical Strategies for Implementing Ethical Information Management and Governance. Kogan Page.





Awarding
Great British
Qualifications

Topic 11 –Data Science Ethical and Privacy Issues

Any Questions?