WHERE IS DATA ETHICS IN POSTSECONDARY DATA SCIENCE COURSES?

Alex Lyford and Katelyn Mei
Department of Mathematics, Middlebury College, Middlebury, VT, USA
alyford@middlebury.edu

The number of postsecondary data science courses, especially at the introductory level, has increased dramatically over the past decade. As the prevalence and complexity of data-related challenges increases, so too does the need to teach students how to handle these issues in an ethically-responsible manner. In this work, we use a combination of course syllabi and course descriptions from 361 data science courses across 81 postsecondary institutions in the United States. We find that fewer than half of all data science courses contain any elements of data ethics, broadly defined. We conclude with examples of the different ways that data ethics is integrated into data science courses across the United States and a vision for data ethics moving forward.

INTRODUCTION AND BACKGROUND

The demand for data science courses in postsecondary education is increasing substantially and is well-documented (e.g., Miller & Hughes, 2017; Salian, 2017; Tate, 2017). Because citizens across a wide variety of roles rely on data to make socially-impactful decisions, the influence of data science on our everyday life is inevitably increasing. With the tools and techniques learned from data science courses, students can collect and analyze many kinds of data at any time and for any purpose. This can pose risks and challenges to society if these tools and techniques are not used with care. As more and more students take courses in data science, their collective impact—whether positive or negative—on society will likewise grow. As such, it is increasingly important that educators instill a sense of moral responsibility and ethical direction in their data science students. Aptly stated by Baumer et al. (2022): "Indifference to ethics in data science is not an informed position" (p. 15).

In this paper, we seek to quantify the prevalence of data ethics in data science courses—broadly construed—at postsecondary institutions in the United States through a combination of publicly-available course descriptions and course syllabi. We discuss the ramifications of data ethics integration in data science courses—or the lack thereof. Finally, based on the syllabi we gathered from data science courses with fully-integrated data ethics content, we provide a list of ways that instructors of data science courses can integrate data ethics deeply into their own courses.

DATA COLLECTION AND DEFINITIONS

To assess the prevalence of data ethics in data science courses, we sought to identify a sufficiently large sample of postsecondary institutions in the United States that offer majors, minors, or significant coursework in data science or a related field (such as data analytics). To achieve this, we used the *U.S. News & World Report* (2022) listings, identifying all schools with data science programs under the *Best Undergraduate Data Science Programs, Top Public Schools,* and *National Liberal Arts Colleges* lists, yielding a total of 81 institutions. Of these schools, 36 offered majors in data science or a related field, 14 offered minors, and the remaining 31 offered neither, although most of these institutions offered certificates, concentrations, or other data-science focused curricular pathways. We then used a combination of web scraping and manual searching through department websites to identify all data science courses at each of the 81 institutions. This yielded a total of 361 courses. We collected the most updated course descriptions of each course and searched for the most recent detailed syllabi available online. We considered a syllabus to be detailed if it contained information about day-by-day or unit-by-unit course content or specific and comprehensive information about the course's learning goals. In total, we obtained detailed syllabi for 98 out of 361 courses.

There is certainly no unanimity in defining data science. This paper does not seek to contribute anything to that discussion. Instead, we attempt to sidestep the issue by taking a conservative approach to defining a data science course. Generally speaking, courses that self-identified as data science courses in any way were included in our study. We also included courses that contained significant components of any of the following: data wrangling, data visualization, statistical modeling, web scraping, text analysis, storytelling with data, and other data-related methods. We excluded courses that were purely mathematical science or statistics courses (e.g., causal inference, regression, or probability theory). We

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org ©2022 ISI/IASE

also excluded domain-specific applied courses and special topics courses. Additionally, we flagged a data science course as *introductory* if both of the following two criteria were true: (a) the course was the first (or only) course in a listed program of study, and (b) the course required at most one prerequisite course (such as introductory statistics or introduction to computer science). Regardless of the previous two criteria, we flagged any course with a name similar to *Introduction to Data Science* as both an introductory and data science course—who are we to judge!?

Next, the two authors of this paper independently read through each of the 361 course descriptions and coded them as either containing a data ethics component or not containing a data ethics component. Courses coded as having an ethics component must include a discussion of at least one of the following:

- Data ethics
- Data privacy
- Emphasis on critical thinking, fairness, equality, or bias
- Impact of data or algorithms on race, gender, or class
- Focuses on social or economic justice using data

Our inter-rater reliability in coding was very strong, with Cohen's Kappa (Cohen, 1960) equal to 0.94. After reconciling differences, we repeated the same procedure using course syllabi. (For course syllabi, we applied the aforementioned rubric by looking through lecture notes, activities, course objectives, assignments, and class schedules.) We found course syllabi by scouring course, department, institution, and instructor websites. We only included detailed syllabi in our analysis—those that listed course learning objectives, day-by-day or unit-by-unit topics and activities, lectures, and notes, etc. Our inter-rater reliability for coding syllabi was likewise strong, with Cohen's Kappa equal to 0.87. Finally, we also coded courses with detailed course syllabi as either having no elements of data ethics, one day of data ethics, one unit of data ethics, or fully-integrated data ethics components. Descriptions of each of the 361 courses, including links to their corresponding syllabi (where applicable), can be found here. Approximately half of these courses are at the introductory level, most of which had no pre-requisites. These courses were taught across a variety of disciplines, and most focused on data techniques such as data visualization, data wrangling, statistical modeling, and computing techniques.

For the remainder of our work, we treat the content of detailed syllabi as the *ground truth* for course content. In other words, we will assume that content described in the lectures, activities, course objectives, etc. of detailed syllabi matches exactly what is covered in the course. Conversely, we will assume that things *not* mentioned in the detailed syllabi are *not* covered in the course. Although we feel that this is a reasonable assumption, we discuss the limitations of this approach at the end of our report.

RESULTS

Table 1 displays the cross-tabulations of data ethics courses by medium and introductory status. In general, data science courses at all levels were unlikely to contain any data ethics components. There were not substantial differences in ethics-related content between introductory and non-introductory courses using either course descriptions or course syllabi. Using detailed course syllabi as the *ground truth* for course content, only 40% of both introductory and non-introductory data science courses were coded as having data ethics. This estimate was 26% solely based on course descriptions.

Table 1. Cross-tabulation of ethics courses and introductory courses by course description and syllabi

	Course Descriptions		Syllabi	
	Introductory	Not Introductory	Introductory	Not Introductory
Ethics	43	49	19	21
No Ethics	100	168	27	31

For many of the courses we examined, the content in online course descriptions matched the content in the corresponding syllabi. Table 2 shows the relationship between the way in which a course's course description was coded and the way its detailed syllabus was coded, where available. In total, we obtained 98 syllabi from the 361 courses. We note that a course is often taught by different instructors across multiple sections and semesters. Our approach assumes that the most recent syllabus reflects the content of the most recent course offering.

Table 2. Agreement and disagreement between course descriptions and syllabi

		Syllabi	
		Ethics	No Ethics
Course Description	Ethics	19	6
	No Ethics	21	52

Based on our findings, only 6 courses whose course descriptions contained elements of data ethics did not actually contain elements of data ethics based on the corresponding course syllabus. Conversely, roughly 50% of courses (21 in total) containing data ethics components did not mention these at all in their course descriptions. Thus, using course descriptions alone provides a conservative *lower bound* for the prevalence of data ethics in data science courses. This result may prove useful for those wishing to generalize the prevalence of course content beyond courses for which syllabi are readily available. For the remainder of our work, we will focus on courses with detailed syllabi because they provide a more comprehensive look at course content.

Using the 40 syllabi from data science courses with ethics components, Table 3 shows the breakdown of how data ethics was integrated into each course. A total of 16 out of the 40 ethics courses (40%) had data ethics content interwoven throughout the course. This included multiple lectures, activities, and/or readings with ethics-related content. Most of these syllabi had data ethics-related statements as part of multiple course objectives or learning outcomes. Nine courses featured a single unit on data ethics, with no mention of data ethics in other course materials. Ten courses apportioned a single day devoted to data ethics. The remaining five courses mentioned data ethics training as some part of their course, but its specific integration was unclear from the materials provided. Of the 19 courses that devoted a single day or unit to data ethics, 12 of these units/days occurred in the last week of class.

Table 3. Integration of ethics into data science course syllabi

Fully Integrated	One Unit	One Day	Other
16 (40%)	9 (22.5%)	10 (25%)	5 (12.5%)

WHAT DOES A FULLY-INTEGRATED DATA ETHICS COURSE LOOK LIKE?

Fully integrating data ethics into data science courses at every level is critical. Limiting discussions of data ethics to a single day (especially the last day of the course!) or a single unit risks students compartmentalizing data ethics as a tangential, non-essential step in the data gathering, cleaning, modeling, and visualizing process. In other words, students who are taught to consider the ethical and moral implications of data *after* learning technical skills may erroneously believe that the technical aspects of data cleaning and the considerations of data ethics should occur mutually exclusively. We believe that fully integrating data ethics into a data science course requires a consistent and intentional discussion of the ethical and moral implications of data-related decisions throughout the course material—through lectures, readings, assignments, etc. This does not require, however, that every activity be centered around data ethics. Instead, students should engage early and often with data ethics-related content to ensure that students understand that data ethics should play an integral role in every step of the data pipeline.

Here, we highlight a prototypical example of data ethics integration into a data science course. Fully integrating data ethics into data science courses requires not only well-defined learning objectives related to data ethics but also a variety of class activities and readings that focus on this objective. Although some of the courses with fully-integrated data ethics components are centered around the topic of data ethics, there are numerous courses that maintain a balance of teaching techniques and discussion of the potential social implications of those techniques. One such example in our data is the course *Responsible Data Science,* offered at New York University (Stoyanovich, 2019). (We note that the authors of this work have no affiliation with New York University—we simply see this course as an exemplar for data ethics integration.)

This second course in data science is structured around the topics of *responsibility, anonymity and privacy, data cleaning, algorithmic fairness, diversity,* and *transparency*. Each topic has corresponding readings, lab tasks, and homework assignments that reinforce these ideas. For example, for the module on anonymity and privacy, students are assigned to read from *The Algorithmic Foundations of Differential Privacy* by Aaron Roth and Cynthia Dwork (2014) and to generate privacy-preserving synthetic datasets to learn about ways in which anonymity can be upheld with minimal loss in the validity and authenticity of the data. In other modules, students might be expected to write a summary report on the assigned readings or a paper on case studies. Integrating the considerations of data ethics into the practice of data science pipelines, these various assignments help students deepen their understanding of each topic, which encourages students to deeply engage with the learning objectives.

Based on the content of the 16 data science courses with fully-integrated data ethics components, we compiled a list of common heuristics for covering various topics in data ethics. Table 4 shows four major topics covered in at least half of these courses, including relevant content and examples.

Table 4. Heuristics for integrating data ethics into data science courses

Topics	Content for Instructors	Examples
Data Privacy	Responsible data sharing, anonymization techniques, limits of anonymization and harms of re-identification	“The Algorithmic Foundations of Differential Privacy” (Dwork & Roth, 2014) Chapter 1,2,3
Data Collection	Ethics of collecting data of human subjects	“Everything We Know About Facebook’s Secret Mood-Manipulation Experiment” (Meyer, 2014)
Data Cleaning	Outlier detection, quantitative and qualitative error detection, documentation of data cleaning procedure	“Data Cleaning: Overview and Emerging Challenges” (Chu et al., 2016)
Data Visualizations	Misleading axes and proportional ink	“How Charts Lie” (Cairo, 2019) “The Visual Display of Quantitative Information” (Tuft2, 2001)
Algorithmic Bias/Fairness	Taxonomy on fairness, bias in the mechanics of data analysis	“Machine Bias” (Angwin et al., 2022) “On the (Im)possibility of Fairness” (Friedler et al., 2016) “The Hidden Biases in Big Data” (Crawford, 2013)

CONCLUSIONS AND LIMITATIONS

In this work, we began by providing a quantification of the prevalence of data ethics content in 361 data science courses at 81 postsecondary institutions in the United States. Across both introductory and non-introductory courses, most data science and related courses do not contain components of data ethics in any part of their course descriptions or syllabi. This is inherently problematic—students' learning tools and techniques of data science at any level should simultaneously consider the many ethical and moral implications of each step in the data gathering, cleaning, visualizing, and modeling process.

Next, we investigated the ways in which data ethics were integrated into courses across curricula. We found that, using only course descriptions, 26% of data science courses contained elements of data ethics. Using detailed course syllabi, only 40% of data science courses contained elements of data ethics. Of the courses whose syllabi contained elements of data ethics, 60% relegated discussions of ethics to a single unit or single day of class.

Finally, we provided a set of heuristics for integrating data science topics such as data privacy, data cleaning, and algorithmic fairness into data science courses. These heuristics, complete with relevant reference material, can help instructors who wish to incorporate these elements of data ethics into their own coursework.

We note that our findings and quantifications of the prevalence of data ethics are almost certainly a lower bound for the true prevalence of ethics-related content in data science courses. (In other words, course descriptions and course syllabi that contain direct references to data ethics topics almost assuredly contain course content related to data ethics, but course descriptions and syllabi that *do not* mention data ethics do not necessarily lack data ethics course content.) Nevertheless, it is clear that many data science courses do not integrate data ethics in any way into their course's learning objectives or course content. We also note that the 81 institutions in our study were chosen for their established programs in data science, so our findings may not be generalizable to institutions without formal data science programs or to institutions outside the United States. We recognize that course descriptions and course syllabi change over time, and many course descriptions may be out of date for the courses they represent.

Our hope is that this work serves both as an indicator of the lack of data ethics training for many students in data science courses and as a primer for how an instructor of a data science course could integrate data ethics-related content into their own courses.

REFERENCES

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. In K. Martin, *Ethics of data and analytics: Concepts and cases* (pp. 254–264). CRC Press. https://doi.org/10.1201/9781003278290

Baumer, B. S., Garcia, R. L., Kim, A. Y., Kinnaird, K. M., & Ott, M. Q. (2022). Integrating data science ethics into an undergraduate major: A case study. *Journal of Statistics and Data Science Education*, *30*(1), 15–28. https://doi.org/10.1080/26939169.2022.2038041

Cairo, A. (2019). *How charts lie: Getting smarter about visual information*. WW Norton & Company.

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016, June). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 international conference on management of data* (pp. 2201–2206). Association for Computing Machinery. https://doi.org/10.1145/2882903.2912574

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Crawford, K. (2013). The hidden biases in big data. *Harvard Business Review, 1*(4). https://hbr.org/2013/04/the-hidden-biases-in-big-data

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), 211–407. https://doi.org/10.1561/0400000042

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). *On the (im)possibility of fairness*. arXiv.1609.07236. https://doi.org/10.48550/arXiv.1609.07236

Meyer, R. (2014, June 28). *Everything we know about Facebook's secret mood-manipulation experiment*. The Atlantic. https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/

Miller, S., & Hughes, D. (2017). *The quant crunch: How the demand for data science skills is disrupting the job market.* Burning Glass Technologies. http://burning-glass.com/wp-content/uploads/The_Quant_Crunch.pdf

Salian, I. (2017, September 5). Universities rush to add data science majors as demand explodes. *San Francisco Chronicle*. https://www.sfchronicle.com/business/article/Universities-rush-to-add-data-science-majors-as-12170047.php

Stoyanovich, J. (2019). *DS-UA 0202: Responsible data science* [Syllabus]. College of Arts and Sciences, New York University. Retrieved July 25, 2022, from https://dataresponsibly.github.io/rds/

Tate, E. (2017, March 15). *Data analytics programs take off*. Inside Higher Ed. https://www.insidehighered.com/digital-learning/article/2017/03/15/data-analytics-programs-taking-colleges

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Graphics Press. https://www.visualizingsociety.com/class/05/notes/vdqch2.pdf

U.S. News & World Report. (2022). *U.S. News & World Report's rankings and advice.* Retrieved June 17, 2022, from https://www.usnews.com/rankings