

ORIGINAL RESEARCH

Open access

Designing an ELSA-Oriented Curriculum for Data Scientists: Integrating Ethical, Legal, and Societal Perspectives into Data Science Education

Saif Al-Hinai^{1*}, Amal Al-Balushi¹, Mohammed Al-Maskari², Sultan Al-Mahruqi³

Abstract

The rapid expansion of artificial intelligence (AI) systems across numerous sectors in recent years has drawn significant attention to the ethical, legal, and societal aspects (ELSA) of these technologies, along with the specific difficulties they introduce. This paper outlines an ELSA curriculum tailored for data scientists. The framework is designed to heighten awareness of ELSA dilemmas in professional practice, establish a shared vocabulary with domain experts to facilitate collaborative problem-solving, and ultimately imbed ELSA principles directly into data science workflows. Rather than being treated as an obstacle or an auxiliary requirement, ELSA ought to be viewed as an essential component of the Data Science Project Lifecycle. The proposed educational model utilizes the CRISP-DM (CRoss-Industry Standard Process for Data Mining) framework as its structural foundation, creating a vertical division composed of modules that align with each CRISP-DM stage. Vertically intersecting these phases is a horizontal division comprising three distinct thematic strands: ethical and societal, legal, and technical rendering knowledge units (KUs). Alongside an exhaustive breakdown of these KUs, this paper examines practical implementation factors, including course duration, delivery formats, assessment methods, and strategies for managing the diverse baseline knowledge and varying requirements of the target student body.

Keywords Artificial intelligence, AI ethics, Education, Data science, ELSA

*Correspondence:

Saif Al-Hinai

saif.alhinai@squ.edu.om

¹ Department of Data Science Education and ELSA Curriculum, College of Engineering, Sultan Qaboos University, Muscat, Oman

² Department of Ethical and Legal Perspectives in Data Science, College of Engineering, University of Nizwa, Nizwa, Oman

³ Department of Societal Implications of AI, College of Engineering, Dhofar University, Salalah, Oman

Introduction

The widespread adoption of data science tools across various industries in recent years has highlighted a range of ethical, legal, and societal aspects (ELSA) associated with data-reliant methodologies. While the nomenclature "data science" was first introduced in 1974, it required nearly three decades to be recognized as an independent academic and professional discipline [1], expanding primarily upon the "core areas of technical work within the discipline of statistics" [2]. Nevertheless, the rapid proliferation of data-centric applications—particularly driven by contemporary AI advancements—has introduced diverse ethical, legal, and societal problems, prompting critics to label these algorithms as "Weapons of Math Destruction" [3]. These concerns encompass a broad spectrum of issues, including breaches of personal data privacy, intellectual property disputes, systemic discrimination against vulnerable demographics and individuals, and ecological degradation (for a detailed overview, [4]).

Several of these challenges have already become the focus of regulatory frameworks. For instance, the European Union established the General Data Protection Regulation (GDPR) [5] to safeguard personal information and, more recently, introduced the AI Act [6]. Conversely, ethical and societal considerations have largely been addressed through recommendations and frameworks drafted by professional, national, international, and supranational entities (for meta-analyses of these guidelines, see [7, 8], while a comprehensive repository is available in [9]).

A recurring recommendation within these frameworks emphasizes the integration of ELSA training for AI engineers and practitioners. Notably, the European Commission's High-Level Expert Group on Artificial Intelligence, in their ethics guidelines for trustworthy AI [10], emphasizes that education and awareness-raising are vital for fostering an ethical mindset among all participants, including software developers (pp. 23–24), and serve as a crucial mechanism for implementing accountability standards (p. 31). Furthermore, cultivating an awareness among engineers that their technological creations are inherently linked to ethical and societal consequences is considered a fundamental prerequisite for building a genuine professional ethos [11]. Because these data-driven models are increasingly deployed to guide choices regarding the distribution of public resources and directly influence civil liberties, arguments have been raised that their creators should undergo formal certification processes akin to medical or legal practitioners, thereby establishing data science as a regulated profession [12, 13].

In this article, we outline a proposed ELSA educational framework for data science practitioners, developed under the auspices of the FAIR Data Spaces initiative [14], which receives financial support from the German Federal Ministry of Education and Research (BMBF). This initiative remains ongoing; the initial iteration of the curriculum was published as an official project output [15], providing a comprehensive breakdown of the educational framework. The second and final iteration is scheduled for publication by the conclusion of 2024.

Background

To properly define the educational prerequisites that an ELSA curriculum must satisfy, we executed a comprehensive review of existing literature—comprising academic papers, industry surveys, and institutional reports—focusing on three core dimensions: a. the demographic profiles and professional perspectives of data scientists regarding ELSA dilemmas, b. corporate demand for competencies tied to ELSA, and c. existing university coursework in this domain (frequently categorized broadly under "ethics," though a closer analysis of syllabus contents demonstrates that they routinely encompass legal and societal dimensions as well). This investigation was carried out during the final quarter of 2021 as part of a published project report mapping the current state of ELSA training [16].

The profile of the data scientist

According to the individuals credited with introducing the phrase, a data scientist represents "a high-ranking professional with the training and curiosity to make discoveries in the world of big data" [17]. This foundational characterization of data scientists and their professional duties contains no reference to ethical or legal dimensions. Conversely, in [18], data science is conceptualized as an intersection of disparate subdisciplines; alongside technical and scientific fields like statistics, machine learning, and systems architecture, it incorporates behavioral and social sciences to address ethics and comprehend human behavior. Furthermore, Ref. [19] expands these technical competencies to encompass stakeholder engagement.

To understand the perspectives of data scientists and their stances on the ELSA dilemmas inherent in their professional activities, we analyzed surveys administered by the Anaconda Python distribution [20] and Kaggle [21], a Google subsidiary and online network for data science and machine learning professionals. These annual surveys predominantly focus on technical topics, drawing voluntary contributions from individual practitioners and corporate bodies globally. Although the participant sample is not strictly controlled, it offers valuable insights into the practitioner demographic. By examining the self-reported profiles of global practitioners in the Anaconda [22–25] and Kaggle assessments [26], we concentrate on two core themes: the educational attainment of data scientists and their perspectives on ELSA issues.

Both surveys indicate that most data scientists hold a higher education degree, with a Master's degree representing the most prevalent qualification. Interestingly, Anaconda's data demonstrates a temporal increase in degreed respondents from 2020 to 2021, whereas Kaggle's consolidated data spanning 2017 to 2021 shows an inverse pattern. The Kaggle analysis attributes this trend to the "increasing availability of online courses and ways to build technical skills and experience, having a degree isn't a prerequisite for getting started in data science" [26].

This discrepancy in findings could stem from variations in the respective survey samples or differing chronological scopes: the Anaconda dataset tracks the years 2020–2023, whereas the Kaggle dataset spans 2017–2021 (Kaggle's 2022 educational metrics were unavailable at the time of this research). One could infer that during the 2017–2021 timeframe, the supply of specialized graduates for data science roles was limited, a dynamic that shifted in subsequent years. Indeed, the World Economic Forum's 2020 Future of Jobs report [27] indicates that the majority of individuals transitioning into AI and data science roles migrate from separate job families (which classify occupations based on daily tasks, skill sets, education, training, and credentials). This transition path is enabled because data science lacks formal professional associations or mandatory certifications, while simultaneously experiencing surging market demand (both the 2020 and 2023 [28] Future of Jobs reports confirm robust demand for data science professionals).

Does the industry demand ELSA skills?

In [29], corporate executives managing IT operations (such as CIOs, CTOs, and digital innovation managers) categorize anonymity, privacy, and ethics among the fundamental domains a data scientist must master. Additionally, a 2018 survey of ACM members [30] revealed that approximately one-third (34%) of industry employers mandate ethical training for applicants, whereas over half (51%) consider it an optional qualification.

Nonetheless, the World Economic Forum's 2023 Future of Jobs report [28] noted that when information and technology service firms ranked ethics as a fundamental skill for employees, only 12% prioritized environmental stewardship, and a mere 7% prioritized global citizenship. Even so, these organizations project that the importance of these competencies will expand in the near term (2023–2027).

This shifting outlook is mirrored in consecutive Anaconda surveys: in 2021, one-third of surveyed organizations stated they had no strategy to ensure fairness, mitigate algorithmic bias, or enhance model explainability, while an equal portion did not know whether such a strategy existed [23]. By 2022, the share of firms employing no safeguards regarding these issues fell to 24%, with "unsure" responses declining to roughly 15% [24]. By 2023, the data indicated that companies were actively recruiting, or planning to recruit, specialized AI ethicists [25].

What do the universities teach?

Because most data scientists possess a university degree, evaluating whether they receive ELSA training during their academic preparation is highly relevant.

According to the previously cited 2018 ACM member survey [30] regarding Computing Competencies for Undergraduate Data Science Curricula, just over half (54%) of data science degree pathways mandated ethics modules, while fewer than 25% of the examined programs actually provided such coursework. Similarly, a 2020 European study investigating ethics education in computer science and adjacent majors revealed that 36% of surveyed institutions offered no ethical training. Among the universities that did provide ethics education, the focus was primarily concentrated within AI-related tracks (64%) and machine learning specializations (51%) [31].

In contrast, Anaconda's industry surveys indicate that only about 20% of students report encountering data science or machine learning ethics within their coursework, while a slightly higher proportion (22–24%) note that algorithmic bias is regularly discussed in their lectures or classes [23, 24].

What the data scientists say about ELSA challenges

Regarding practitioner perspectives on ELSA concerns, Anaconda's findings reveal that data scientists view the societal consequences of algorithmic and data bias as the most critical challenge facing AI/ML, followed closely by threats to individual privacy [22–24]. Despite these concerns, when evaluating internal skill gaps, only 15% of respondents identified ethics as a competency currently missing among the data professionals in their organizations [23].

The 2023 survey focused specifically on generative AI applications. Here, the largest share of participants (39%) emphasized the necessity of transparency and interpretability in AI systems, followed by concerns over algorithmic bias and fairness (27%), and the challenge of balancing copyright and intellectual property protections against technological innovation (15%) [25].

Conclusions

The systemic landscape of ELSA training for data scientists—mapped through our evaluation of existing academic courses and datasets detailing practitioner profiles, industry requirements, and self-perceptions—leads to the following conclusions:

- Data scientists originate from both computer science and statistical academic pathways, with an increasing cohort arriving from external fields like social sciences, finance, and business; furthermore, a notable segment enters the data science profession via non-traditional or quasi-academic routes.
- Academically trained professionals have not systematically undertaken ELSA-specific courses, whether as compulsory components or electives.
- Modern industry values but does not strictly mandate ELSA competencies when hiring data science personnel.
- A rising proportion of data scientists acknowledge the necessity for heightened awareness and actionable responses regarding the ELSA complications present in their professional work.

Curriculum Proposal

Following our foundational analysis aimed at identifying the educational gaps a data science ELSA curriculum should fulfill, we formulated the strategic framework summarized in this section. Our development process prioritized the following core considerations:

- Data science initiatives are inherently multidimensional across several vectors: a. their domains of application, where an analysis of academic literature reveals a minimum of 20 distinct involved disciplines [1]; b. execution, which relies on the collaborative input of diverse specialists spanning multiple fields (such as computer scientists, engineers, statisticians, and visualization experts); and c. broader societal implications, which necessitate direct engagement with external stakeholders, including legal counsel, ethicists, and public or community representatives (such as patient cohorts in medical research).
- ELSA dilemmas emerge continuously across the entire data science lifecycle. While certain issues span multiple iterative stages, others remain localized to specific tasks. Similarly, while some challenges impact all project participants uniformly, others affect specific team members to varying degrees. Regardless of scope, these factors must be handled as an invariant, structural component of the engineering pipeline rather than an optional or peripheral consideration.

We addressed these priorities by embedding flexibility into the fundamental design and educational vision of the program, opting for a modular architecture capable of adapting to diverse educational contexts. The structural benchmark of this curriculum is the knowledge unit (KU)—defined as a distinct, concise learning block dedicated to a specific ELSA topic.

These KUs are organized within a matrix formed by a horizontal axis representing thematic disciplines and a vertical axis representing the operational phases of a data science project.

The subsequent sections outline our primary educational goals and curricular vision, followed by a detailed mapping of the structural components and individual modules.

General objectives and vision

Reflecting on the professional responsibilities and ELSA obstacles encountered by active practitioners, this curriculum is built around three primary educational benchmarks. Specifically, the curriculum aims to train data scientists to:

1. Identify ethical, legal, and societal dimensions tied directly to their technical output (awareness);
2. Establish a shared technical and conceptual vocabulary with domain specialists to foster cooperative problem-solving (communication ability);
3. Embed ELSA considerations directly into the technical development lifecycle, viewing these frameworks not as bureaucratic constraints or superfluous documentation, but as essential pillars of standard engineering practices (professional mentality building).

Curriculum structure and overview

As previously noted, the curricular matrix consists of a vertical dimension (modules) and a horizontal dimension (strands) that cross-reference to position each knowledge unit. A KU serves as the foundational element of the program, operating as an explicit instructional block focused on a designated ELSA concept.

The horizontal division is derived from a. a thematic synthesis of current university course catalogs [16] and b. our first two instructional goals (awareness and communication ability). To meet these benchmarks, engineering students must grasp core theoretical principles from external fields—namely jurisprudence, moral philosophy, and the social sciences—while learning to apply specific technical standards, frameworks, and auditing methodologies. Consequently, we organized the material into three overarching thematic areas:

- **Ethical and Societal Knowledge Units:** These blocks explore the ethical boundaries of computational systems. Topics range from aligning software design with community priorities and preventing demographic discrimination or systemic bias to mitigating the environmental footprint of large-scale infrastructure and establishing clear frameworks for executive accountability and liability.
- **Legal Knowledge Units:** These modules equip technical teams to navigate the regulatory landscapes governing their day-to-day operations. Instruction concentrates on data privacy mandates, intellectual property boundaries, copyright constraints, and essential legal nomenclature.
- **Technical Renderings Knowledge Units:** These units translate abstract legal and ethical obligations into functional engineering requirements. Focus areas include technical privacy controls, algorithmic bias detection and correction techniques, and the integration of fairness, validity, and explainability metrics into model validation. They also cover production-level system monitoring and deployment strategies outside sandboxed environments. These units emphasize practical case studies to illustrate how software tools solve these issues, rather than teaching baseline engineering skills, which a. are typically covered in core technical coursework or professional practice and b. vary significantly across specialized application domains.

The vertical division (modules) fulfills our third objective: integrating ELSA compliance smoothly into active engineering workflows. To achieve this, we adopt the CRISP-DM (CRoss-Industry Standard Process for Data Mining) model. CRISP-DM is an open-source, vendor-neutral, and application-independent data mining framework widely utilized as an industry benchmark. The framework divides a project into six distinct phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment, mapping out a structured lifecycle for executing data-driven initiatives [32].

While alternative framework models like KDD (Knowledge Discovery in Databases) [33] or SEMMA (Sample, Explore, Modify, Model, Assess) [34] could serve a similar structural purpose due to the direct functional mapping that exists between their respective steps [35], CRISP-DM was selected for three specific reasons: a. its widespread adoption and visibility across commercial data science teams; b. its origin in industrial practice, making it highly applicable to working professionals; and c. its established history as a foundational model for analyzing legal liabilities, mapping ethical issues, and engineering compliance toolkits [36–38].

Figure 1 outlines this intersecting matrix. The ethical and societal track is highlighted in grey, the legal track in light blue, and the technical rendering track for ELSA solutions in deep blue.

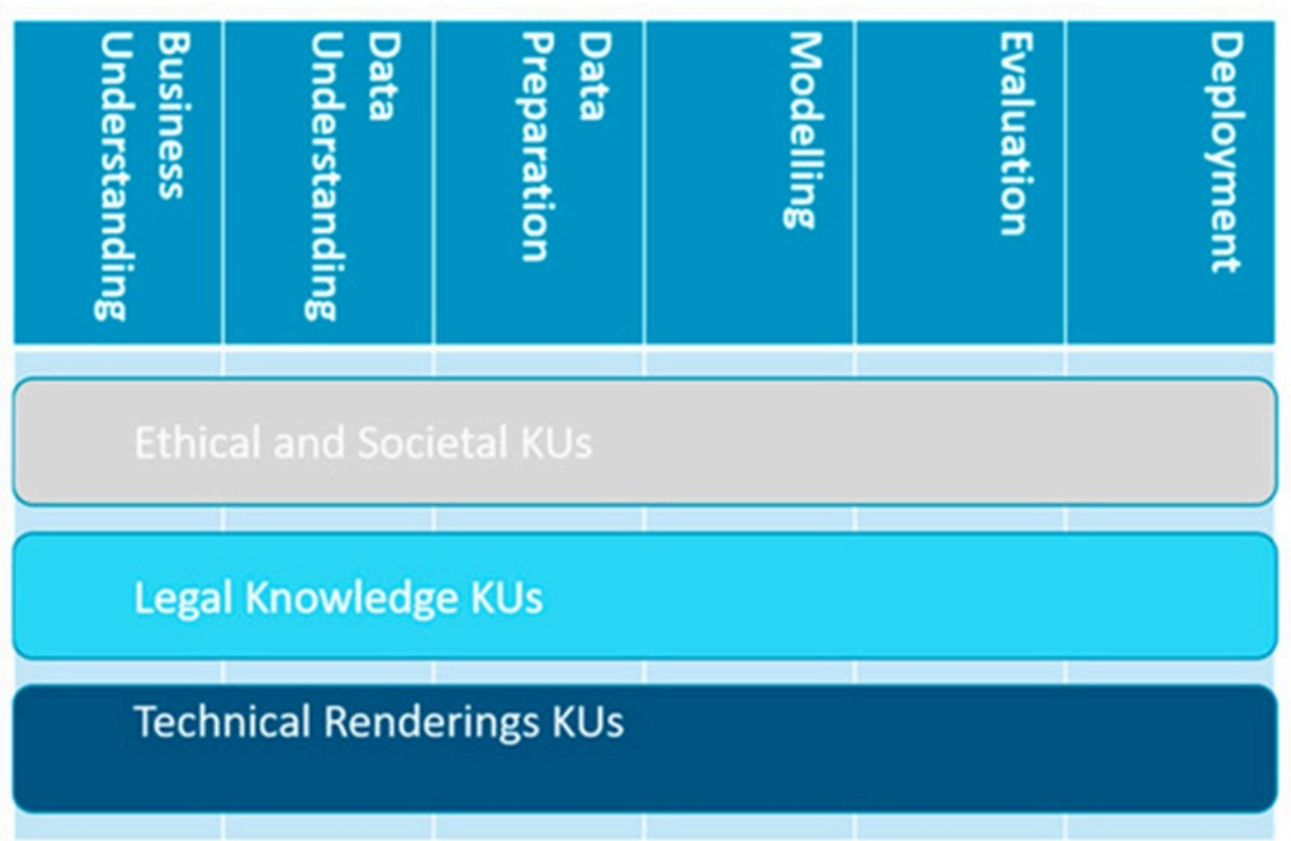


Figure 1. The horizontal–vertical structure of the curriculum. With grey, we denote the ethical/societal strand, with light blue, the legal strand, and with deep blue, the technical renderings of the ELSA challenges. The vertical structure is expressed as modules that correspond to the CRISP-DM phases, while the horizontal structure comprises knowledge units belonging to three strands that run through phases.

ELSA Curriculum KUs

This section provides a structural review of the designated KUs, outlining the core competencies without mapping every edge case. For each distinct module—aligned with the standard CRISP-DM phases detailed in [32]—we present a brief summary of its respective KUs.

Module I: Business understanding

This baseline phase centers on defining project goals from a corporate or institutional viewpoint, translating those high-level ambitions into concrete data science objectives, and drafting the initial technical roadmap.

The corresponding KUs focus on rigorous stakeholder mapping; the integration of localized values through professional ethical codes, national regulations, and international operational standards; and the application of "by design" frameworks for privacy and data compliance. Additionally, this module explores organizational culture—specifically how corporate workflows and auditing procedures incentivize or support ethical decision-making—alongside foundational legal concepts, including jurisdictional boundaries, digital governance, statutory hierarchies (such as national versus EU law alignment), and liability management tools like system audits, impact assessments, and risk mitigation strategies.

Module II: Data understanding

This phase targets initial ingestion pipelines, exploratory data analysis, and validation checks to verify data integrity and isolate anomalies early in the cycle.

This instructional module shares overlapping themes with the subsequent phase (data preparation). In specific custom applications of this curriculum, these two segments can be condensed into a single learning track—a structural choice reflected in the data-centric ethical challenges categorization proposed by [36], which informed our overall curriculum design. Nevertheless, we maintain these as distinct modules due to three primary factors: a. critical data ingestion and collection risks must be evaluated before any downstream preprocessing occurs; b. early findings here frequently reshape the initial problem constraints established during the business understanding phase; and c. ingestion and preprocessing are often decoupled in modern industrial settings, meaning data scientists frequently inherit static repositories without having participated in the original collection process.

Consequently, depending on how this curriculum is deployed, KUs covering shared data themes can be moved or distributed across either module based on time constraints and institutional goals.

The specific KUs in this block cover data protection regulations, intellectual property considerations for source material, early-stage data bias detection and mitigation, synthetic data generation, the prevention of offshore ethics dumping, and structured metadata documentation practices.

Module III: Data preparation

This stage centers on constructing the definitive dataset intended for model ingestion, a pipeline encompassing the selection, purification, quality assessment, integration, and standardization of the gathered information.

Although the primary subject areas mirror those of the preceding stage—and the KUs retain identical designations—the pedagogical focus shifts to differentiate the underlying issues. Consequently, depending on the specific educational context, these KUs can be delivered either as standalone components or integrated into a combined framework to prevent redundant instruction. The KUs within this module explore data protection mandates (utilizing a designated regulatory standard as a framework for practical case analysis) and intellectual property rights, specifically examining how training datasets vary by licensing terms within the context of generative AI systems. Furthermore, this stage investigates pre-processing data challenges ranging from label annotation and data cleansing to synthetic data generation, feature selection, and proxy determination.

Students also confront diverse forms of algorithmic bias—specifically representation, measurement, and aggregation bias—alongside broader data validity concerns. For instance, the module highlights how data annotation outcomes diverge when executed through crowdsourcing networks, domain experts, or automated machine learning pipelines. Societal dynamics, such as the ethical implications of employing low-cost labor for data labeling by offshoring operations to the Global South, are critically examined. Finally, dataset documentation practices are advanced here as a direct continuation of the documentation KU established in the prior module.

Module IV: Modelling

This phase focuses on selecting algorithmic architectures and calibrating hyperparameters to align with specific problem definitions and data constraints.

The KUs in this block analyze model-level bias and corresponding mitigation methods, transparency, and explainability frameworks, alongside the environmental footprint associated with computational training. Furthermore, intellectual property concerns are explored regarding the algorithms themselves and their resulting outputs, including the commercial deployment of pre-trained models, trade secrets, ownership of trained weights, and automated algorithmic outcomes. These topics are synthesized through standardized model documentation methodologies.

Module V: Evaluation

This process entails auditing the architectural construction of the model and verifying whether its functional performance satisfies the strategic benchmarks established during the initial business understanding phase.

The KUs assigned to this stage evaluate model performance using metrics that extend beyond standard accuracy, integrating considerations such as algorithmic fairness, resource allocation efficiency, explainability, systemic trustworthiness (robustness outside isolated experimental environments), and whether the chosen features and proxies genuinely resolve the underlying problem despite yielding mathematically accurate results. Particular emphasis is placed on fairness, given that the verification phase delivers the first actionable outputs for assessing potential bias and discriminatory patterns. Additionally, the curriculum addresses different formal taxonomies of fairness, its corresponding statutory provisions, human cognitive perceptions of equity and how they influence product adoption, and theoretical insights from external fields (such as moral philosophy) that extend beyond purely technical interventions. This block concludes with a model documentation KU that builds directly upon the identically named unit from the previous section.

Module VI: Deployment

This final operational phase encompasses the integration, continuous monitoring, and long-term maintenance of the computational system. While the engineering team's clients frequently manage the actual deployment steps, it remains vital for those clients to understand the operational realities and constraints of the deployed system. As established by the authors of [36], the ethical obligations of a data scientist do not conclude upon project delivery. Practitioners bear a professional responsibility to articulate their design choices and subsequent ramifications using accessible language tailored for non-technical stakeholders, such as corporate executives.

Accordingly, a specific KU addresses deployment constraints regarding both known vulnerabilities and actual functional boundaries, while analyzing the chosen degree of automation and the potential consequences of system failures or adverse outcomes. A specialized KU evaluates visualization bias and prevalent graphical misinterpretations; data scientists often lack familiarity with these issues because they are traditionally categorized as user experience (UX)

challenges. This unit aims to bridge that gap, enabling data scientists to collaborate effectively with UX designers to minimize data misinterpretation. Lastly, a KU covers systemic accountability and enforcement mechanisms, serving as a functional counterpart to several topics introduced in Module I (Business Understanding). These include auditing standards, corporate culture, institutional processes, individual liability boundaries, and the arguments surrounding professional certification pathways for data scientists in tandem with ethical codes and formal professionalization.

Figure 2 provides a more comprehensive, though non-exhaustive, schematic mapping of the curriculum components. Each individual node may encompass multiple distinct knowledge units; specific concepts handled within each module are visually detailed in the graphic. In addition to the three core horizontal strands, a dedicated documentation knowledge unit has been integrated across all lifecycle phases, highlighting the mandatory practice of logging every action taken throughout the project. The exact execution of this process varies by lifecycle stage, as this framework does not mandate a singular documentation or auditing tool.

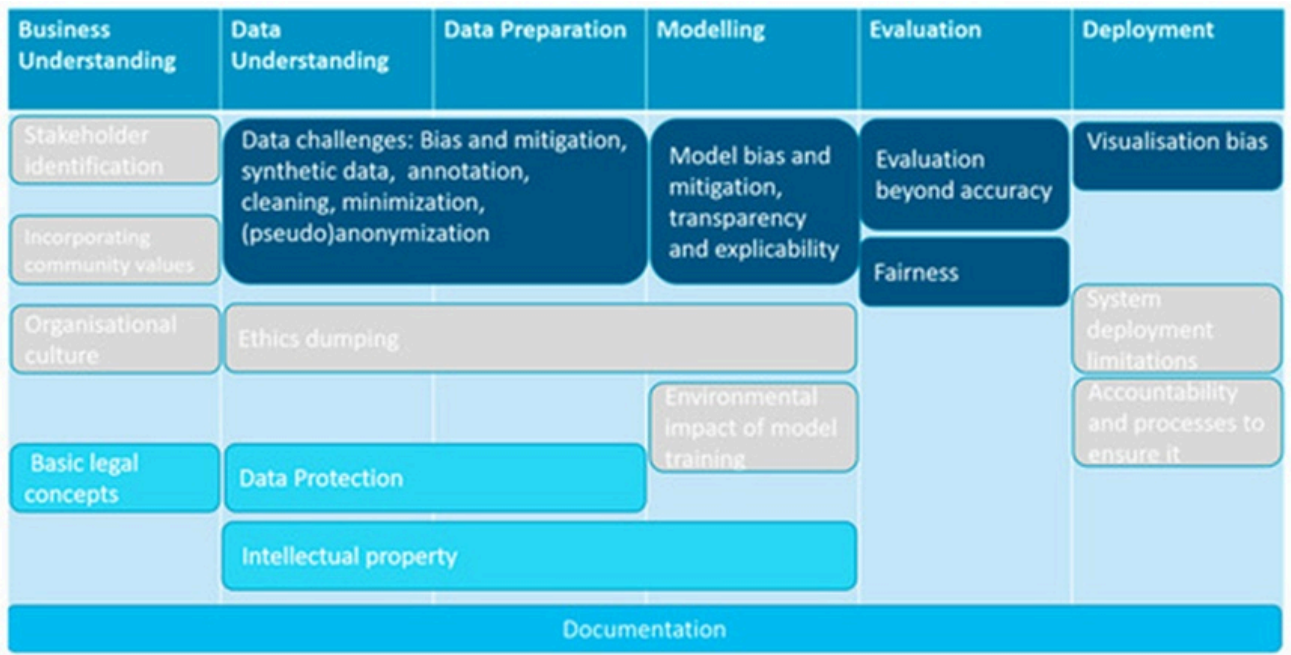


Figure 2. Detailed presentation of the curriculum. Not all KUs are represented, rather the main subjects that are addressed. As in **Figure 1**, with grey, we denote the ethical/societal strand, with light blue, the legal strand, and with deep blue, the technical renderings of the ELSA challenges.

Discussion—Implementation Strategies and Limitations

The structural ELSA curriculum outlined above represents an iterative framework currently being developed as a functional demonstrator within a specific research initiative (FAIR Data Spaces [14]). Consequently, its design is bound by the functional constraints and objectives established by that project; specifically, the initiative aims to interface academic research centers with industrial enterprises under the governance of the FAIR Principles (ensuring data assets remain findable, accessible, interoperable, and reusable).

As a result, this curriculum is explicitly tailored for active data science practitioners who possess foundational field experience executing data-driven initiatives.

The background knowledge and professional requirements of the target cohort are also critical variables for determining the optimal duration and educational delivery format of the program. Data science professionals emerge from highly diverse educational and vocational disciplines, occupying varied organizational roles that extend from software developers writing isolated source code to system architects and technical project managers. Furthermore, different fields of application dictate distinct levels and styles of ELSA literacy; for example, engineering a commercial music recommendation system presents fundamentally different concerns than architecting a clinical healthcare pipeline.

In its current configuration, the curriculum targets junior-level data scientists, offering an overarching structural view of the ELSA friction points they may navigate professionally, thereby bridging the training gaps left by traditional university programs. For this cohort, an intensive, single-week summer school model covering the complete thematic matrix in an introductory manner offers the most efficient delivery mechanism.

Alternatively, the curriculum can be adapted to align directly with the specific workplace responsibilities of the cohorts, the specialized demands of their industry domain, or both. Under this strategy, the most practical implementation method involves expanding or contracting designated KUs to address the specific needs of profiles like technical project managers or software engineers. A parallel adaptation strategy involves tailoring KU case studies to specialized application domains, such as natural language processing (NLP), computer vision, or clinical medical informatics. For these targeted tracking options, an advanced workshop or consecutive seminar series focusing on narrow technical issues represents the ideal delivery format.

Regarding instructional delivery methods and classroom dynamics, our initial assessment of the current ELSA educational landscape indicated that the most prevalent modalities involve standard institutional lectures and guest presentations (sourced from both academia and industry), alongside case-based or problem-based instructional models. These are frequently accompanied by structured debates, peer reflection, group discussions, and role-playing exercises, while interactive, practical activities and real-world case reviews emerged as the most successful pedagogical interventions [16].

The systematic integration of case studies is central to this proposal. As previously established, this ELSA curriculum is being built within an active, multidisciplinary research project (FAIR Data Spaces) that pairs legal scholars and ethicists directly with software engineers to resolve challenges surrounding data privacy, intellectual property rights, and informed consent protocols. Crucially, the parent project features three active technical demonstrators—focusing on ecological biodiversity, data validation and quality control, and federated analytics platforms for clinical health data—all of which serve as foundational use cases within the curriculum. A preliminary model for integrating these demonstrators is outlined in [15], while a comprehensive operational breakdown will be published alongside the definitive curriculum release at the conclusion of 2024.

As noted, this ELSA curriculum remains a work in progress. Ongoing validation of its thematic adequacy is conducted via feedback loops generated through participatory workshops, targeted surveys, and the publication of official project milestones; the gathered evaluations will directly inform the second, definitive iteration of the curriculum. Ultimately, determining the true instructional efficacy of the program requires launching a pilot educational initiative, paired with structured post-course evaluations consisting of participant surveys, instructor focus groups, and qualitative interviews. However, executing such a pilot falls outside the formal scope of the FAIR Data Spaces project.

Conclusion

In this paper, we introduced a structured framework for an ELSA curriculum tailored for data science professionals. In summary, we emphasize the following core developments:

- We constructed a professional profile of the contemporary data scientist, balancing the idealized, multi-talented interdisciplinary practitioner described in academic literature against empirical workforce data and survey metrics

showing a cohort arriving from widely divergent educational and vocational pathways. This profile demonstrates that while institutional guidelines, corporate entities, and practitioners themselves highly value ELSA fluencies, an operational knowledge gap persists among active professionals regarding how to handle these challenges.

- We developed a structured ELSA curriculum that cross-references knowledge units across three core areas (ethical, legal, and technical) within modules aligned with the six traditional phases of the industry-standard CRISP-DM framework.
- The fundamental goals of the curriculum are to heighten awareness of ELSA friction points within data-driven software deployment; improve communication fluencies by establishing a shared technical and conceptual vocabulary with domain specialists (such as legal scholars) to ensure effective collaboration; and cultivate a professional engineering mindset that treats ELSA evaluation as a non-negotiable, systemic pillar embedded throughout the data science project lifecycle.
- Deploying this curriculum effectively requires a major commitment of organizational resources, given the necessary instructional time and its inherently interdisciplinary scope. We propose an adaptable deployment strategy that permits the expansion, contraction, or blending of specific KUs across modules to match the organizational roles and industry domains of the participants. The depth of instruction can be modulated: a holistic, introductory program is recommended for junior professionals, whereas a specialized, deep-dive format focusing on specific operational areas is better suited for senior practitioners.
- Finally, we made the foundational version of this curriculum publicly available and invite critique, feedback, and optimization suggestions from the broader community. Toward this end, we have conducted an initial series of interactive workshops and are gathering targeted stakeholder feedback through a structured survey instrument. These contributions will directly shape the second iteration of the curriculum. Ultimately, the definitive benchmark for evaluating this framework will be its practical execution and subsequent assessment through the direct experiences of both instructors and students.

Acknowledgements

None

Conflict of interest

None

Financial support

This research was partially funded by the German Federal Ministry of Education and Research (BMBF) FAIR Data Spaces project by grant number FAIRDS15.

Ethics statement

None

Received: 23 Jan 2023

Revised: 18 Mar 2023

Accepted: 13 May 2023

Published online: 15 September 2023

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Emmert-Streib F, Dehmer M. Defining data science by a data-driven quantification of the community. *Mach Learn Knowl Extr*. 2019;1:235-51.
- Cleveland WS. Data science: an action plan for expanding the technical areas of the field of statistics. *Int Stat Rev*. 2001;69:21-6.
- Ge R, Kakade SM, Kidambi R, Netrapalli P. The step decay schedule: a near optimal, geometrically decaying learning rate procedure for least squares. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*; 2019; Vancouver, BC.
- Crawford K. *Atlas of AI: power, politics, and the planetary costs of artificial intelligence*. New Haven (CT): Yale University Press; 2021.
- European Parliament, Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). 2016 [cited 2024 Mar 26]. Available from: <http://data.europa.eu/eli/reg/2016/679/oj/eng>
- European Commission, Directorate-General for Communications Networks, Content and Technology. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Brussels: European Commission; 2021 [cited 2024 Mar 26]. Available from: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>
- Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. 2019;1:389-99.
- Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M. *Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI*. Cambridge (MA): Berkman Klein Center for Internet & Society; 2020.
- AlgorithmWatch. *AI Ethics Guidelines Global Inventory* [Internet]. [cited 2021 Feb 10]. Available from: <https://inventory.algorithmwatch.org>
- High-Level Expert Group on Artificial Intelligence. *Ethics guidelines for trustworthy AI*. Brussels: European Commission; 2019.
- Borenstein J, Howard A. Emerging challenges in AI and the need for AI ethics education. *AI Ethics*. 2021;1:61-5.
- Garzcarek U, Steuer D. Approaching ethical guidelines for data scientists. *arXiv [Preprint]*. 2019. <https://doi.org/10.48550/arXiv.1901.04824>.
- Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell*. 2019;1:501-7.
- Nationale Forschungsdateninfrastruktur. *FAIR Data Spaces* [Internet]. [cited 2024 Feb 14]. Available from: <https://www.nfdi.de/fair-data-spaces/>
- Christoforaki M. *ELSA training curriculum for data scientists: version 1.0*. Cologne: Universität zu Köln; 2023.
- Christoforaki M. *ELSA training for data scientists: describing the landscape*. Cologne: Universität zu Köln; 2021.

Davenport TH, Patil DJ. Data scientist: the sexiest job of the 21st century. *Harv Bus Rev*. 2012;90:70-6.

van der Aalst WMP. Data scientist: the engineer of the future. In: Mertins K, Bénaben F, Poler R, Bourrières JP, editors. *Enterprise interoperability VI*. Cham: Springer International Publishing; 2014. p. 13-26.

Luna-Reyes LF. The search for the data scientist: creating value from data. *ACM SIGCAS Comput Soc*. 2018;47:12-6.

Anaconda. About Anaconda [Internet]. [cited 2023 Dec 14]. Available from: <https://www.anaconda.com/about-us>

Kaggle. Kaggle: your machine learning and data science community [Internet]. [cited 2023 Dec 14]. Available from: <https://www.kaggle.com/>

Anaconda. State of data science 2020 [Internet]. [cited 2024 Mar 26]. Available from: <https://www.anaconda.com/resources/whitepapers/state-of-data-science-2020>

Anaconda. State of data science 2021 [Internet]. [cited 2024 Mar 26]. Available from: <https://www.anaconda.com/resources/whitepapers/state-of-data-science-2021>

Anaconda. State of data science report 2022 [Internet]. [cited 2024 Mar 26]. Available from: <https://www.anaconda.com/resources/whitepapers/state-of-data-science-report-2022>

Anaconda. State of data science report 2023 [Internet]. [cited 2024 Mar 26]. Available from: <https://www.anaconda.com/state-of-data-science-report-2023>

Kaggle. Kaggle's state of machine learning and data science 2021 [Internet]. 2021 [cited 2024 Mar 26]. Available from: <https://www.kaggle.com/kaggle-survey-2021>

Zahidi S, Ratcheva V, Hingel G, Brown S. The future of jobs report 2020. Geneva: World Economic Forum; 2020.

Di Battista A, Grayling S, Hasselaar E. The future of jobs report 2023. Geneva: World Economic Forum; 2023.

Mikalef P, Giannakos M, Pappas I, Krogstie J. The human side of big data: understanding the skills of the data scientist in education and industry. In: *Proceedings of the 2018 IEEE Global Engineering Education Conference (EDUCON)*; 2018; Santa Cruz de Tenerife, Spain. p. 503-12.

Danyluk A. Paul Leidig computing competencies for undergraduate data science curricula: ACM Data Science Task Force. New York: ACM; 2021.

Stavarakakis I, Gordon D, Tierney B, Becevel A, Murphy E, Dodig-Crnkovic G, et al. The teaching of computer ethics on computer science and related degree programmes: a European survey. *Int J Ethics Educ*. 2022;7:101-29.

Shearer C. The CRISP-DM model: the new blueprint for data mining. *J Data Warehouse*. 2000;5:13-22.

Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI Mag*. 1996;17:37-54.

SAS Institute Inc. SAS Enterprise Miner: SEMMA [Internet]. [cited 2024 Feb 14]. Available from: <https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jn8bbj1a2.htm>

Azevedo A, Santos MF. KDD, SEMMA and CRISP-DM: a parallel overview. In: Weghorn H, Abraham A, editors. *Proceedings of the IADIS International Conference on Applied Computing*; 2008; Amsterdam, The Netherlands. p. 182-5. Available from: <https://www.iadisportal.org/digital-library/kdd-semma-and-crisp-dm-a-parallel-overview>

Saltz JS, Dewar N. Data science ethical considerations: a systematic literature review and proposed project framework. *Ethics Inf Technol*. 2019;21:197-208.

Rochel J, Évéquoz F. Getting into the engine room: a blueprint to investigate the shadowy steps of AI ethics. *AI Soc.* 2021;36:609-22.

Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci Eng Ethics.* 2020;26:2141-68.