Article

Towards an ELSA Curriculum for Data Scientists

Maria Christoforaki [1,*] and Oya Deniz Beyan [1,2]

[1] Institute for Biomedical Informatics, Faculty of Medicine and University Hospital Cologne, University of Cologne, 50674 Cologne, Germany; oya.beyan@uni-koeln.de
[2] Fraunhofer Institute for Applied Information Technology FIT, Schloss Birlinghoven, 53757 St. Augustin, Germany
* Correspondence: maria.christoforaki@uk-koeln.de

Abstract: The use of artificial intelligence (AI) applications in a growing number of domains in recent years has put into focus the ethical, legal, and societal aspects (ELSA) of these technologies and the relevant challenges they pose. In this paper, we propose an ELSA curriculum for data scientists aiming to raise awareness about ELSA challenges in their work, provide them with a common language with the relevant domain experts in order to cooperate to find appropriate solutions, and finally, incorporate ELSA in the data science workflow. ELSA should not be seen as an impediment or a superfluous artefact but rather as an integral part of the Data Science Project Lifecycle. The proposed curriculum uses the CRISP-DM (CRoss-Industry Standard Process for Data Mining) model as a backbone to define a vertical partition expressed in modules corresponding to the CRISP-DM phases. The horizontal partition includes knowledge units belonging to three strands that run through the phases, namely ethical and societal, legal and technical rendering knowledge units (KUs). In addition to the detailed description of the aforementioned KUs, we also discuss their implementation, issues such as duration, form, and evaluation of participants, as well as the variance of the knowledge level and needs of the target audience.

Keywords: AI ethics; data science; artificial intelligence; ELSA; education

check for updates

Citation: Christoforaki, M.; Beyan, O.D. Towards an ELSA Curriculum for Data Scientists. *AI* **2024**, *5*, 504–515. https://doi.org/10.3390/ai5020025

Academic Editors: Pablo Rivas, Gissella Bejarano, Miguel Molina-Solana and Gianni D'Angelo

Received: 22 December 2023
Revised: 19 February 2024
Accepted: 27 March 2024
Published: 11 April 2024

1. Introduction

The ubiquity of data science applications in recent years has put into the spotlight a multitude of ethical, legal, and societal aspects (ELSA) regarding data-driven methods.

The term "data science" appeared in 1974 but took almost 30 years to be identified as a discipline of its own [1], specifically as an enlargement of "the major areas of technical work of the field of statistics" [2]. However, the proliferation of data-driven applications (especially with the use of AI in recent years) has given rise to a variety of ethical, legal, and societal challenges that led these kinds of applications to be characterised as "Weapons of Math Destruction" [3]. These challenges range from personal data privacy infringement and intellectual property issues to discrimination against vulnerable groups and individuals and environmental issues, among others (for a comprehensive presentation, see [4]).

Some of the above-mentioned challenges have been the subject of regulation, for example, in the European Union, the General Data Protection Regulation (GDRP) [5] for the protection of personal data and, more recently, the AI Act [6].

Ethical and societal aspects, on the other hand, have been the subject of recommendations and guidelines by professional, national, international, and supranational organisations and institutions (for guideline reviews, see [7,8], while an inventory can be found in [9]).

Part of the guidelines' suggestions refers to ELSA education for AI practitioners. Specifically, the European Commission's high-level expert group on artificial intelligence, in their ethics guidelines for trustworthy AI [10], highlight the importance of education and awareness raising both to promote an ethical mindset for all stakeholders, including developers (pp. 23–24), as well as a means to develop accountability practices (p. 31).

Additionally, the need to make developers aware that the technology they are building is intertwined with ethical and societal dimensions is viewed as a first step in establishing an authentic professional mindset [11]. Moreover, since data-driven systems are used to support decisions regarding access to social goods and have an impact on citizen's rights, it may be the case that their developers should be required to be certified like doctors or lawyers, and data scientists should be defined as a formal profession [12,13].

In this paper, we present our suggestion for an ELSA curriculum for data scientists, which is being developed in the framework of the FAIR Data Spaces project [14], funded by the German Federal Ministry of Education and Research (BMBF). It is a work in progress; we published the first version of the curriculum as a project deliverable [15] (see also Supplementary Materials), where a more extensive presentation of the proposed curriculum can be found. The second and final version is due to be released at the end of 2024.

The rest of the paper is structured as follows: in Section 2, we paint with broad strokes the profile of the data scientist as it is presented in the bibliography and emerges in surveys. Using the same sources, we also report any ELSA training they might have received during their studies, the industry attitude towards ELSA, as well as their own opinions on the importance of relevant issues. In Section 3, we present our curriculum proposal, and in Section 4, we discuss the limitations and challenges of the curriculum implementation. Finally, we close with Section 5 where we present our conclusions.

2. Background

In order to identify the needs an ELSA curriculum should address, we conducted a literature review (which, apart from scientific publications, included surveys and reports by international organisations) regarding the following: a. the profile and the attitudes of the data scientists towards ELSA issues, b. the industry demands regarding ELSA skills, and c. the relevant courses provided by universities (usually referred to as "ethics"; however, a more detailed examination of their topics reveals that they also include legal and societal issues). This work took place during the last quarter of 2021 and was part of a published project deliverable describing the ELSA training landscape [16] (see also Supplementary Materials).

2.1. The Profile of the Data Scientist

According to the individuals who claimed to have coined the term, a data scientist is "a high-ranking professional with the training and curiosity to make discoveries in the world of big data" [17]. In this description of the data scientist and their work, no mention of ethical or legal aspects is made. In [18], on the other hand, the authors describe data science as the intersection of different subdisciplines, which, apart from the technical and scientific domains such as statistics, machine learning, and system design, includes behavioural and social sciences for ethics and understanding human behaviour, while Ref. [19] adds to the technical skills of stakeholder involvement.

To gain an insight into the opinions of data scientists themselves and their attitudes about ELSA challenges of their work, we turned to the surveys conducted by Anaconda Python distributor [20] and Kaggle [21], the Google subsidiary and an online community of data scientists and machine learning practitioners. These surveys are conducted yearly and mostly contain questions about technical issues. Participants (individuals and organisations) come from all over the world on a voluntary basis. Undoubtedly, the sample is not rigorously defined. However, it gives us an insight into the profile of the practitioners. Looking at the profile of the data scientist as is self-reported by practitioners from all over the world, in Anaconda [22–25] and Kaggle surveys [26], we focus on two issues: the education level of data scientists and their attitude towards ELSA challenges.

Both of the surveys describe the majority of data scientists as having some tertiary education degree (Master's being the most common). Notably, in the Anaconda surveys, the number of responders with degrees has risen over the years (2020–2021), while in the Kaggle summary of the years 2017–2021, the opposite is true. The survey attributes this 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].

The difference in the findings may be attributed to the variance of the samples between the two surveys or to the fact that the time period covered is different: The Anaconda survey covers the years 2020–2023, while the Kaggle one covers the years 2017–2021 (the Kaggle 2022 survey results regarding the education level of the participants were yet not available at the time of writing). One might hypothesise that in the early years of 2017–2021, there were not many graduates available to work as data scientists, but that changed in the following years. In fact, according to the World Economic Forum’s 2020 Future of Jobs report [27], most of the people transitioning to AI and data science occupations originate from a different job family (job families are groups of occupations based upon work performed, skills, education, training, and credentials); this may be due to the fact that data science as a profession is not bound by professional associations, certifications, etc., and is also in high demand (both the 2020 and 2023 [28] Future of Jobs reports show high demand for data science practitioners).

2.2. Does the Industry Demand ELSA Skills?

In [29], industry executives who have held leadership roles in IT business operations (e.g., CIO, CTO, and digital innovation manager) include anonymity, privacy, and ethics in core areas in which a data scientist must be knowledgeable. Furthermore, in a 2018 ACM members survey [30], one-third (34%) of the industry representatives require experience in ethics from their prospective employees, while half of them (51%) responded that this is elective.

Nevertheless, in the World Economic Forum’s 2023 Future of Jobs report [28], when information and technology services organisations were asked to rank ethics as a core skill for their workers, only 12% did so for environmental stewardship and 7% for global citizenship. However, they see it as increasing in significance for the near future (2023–2027).

The latter is also reflected in the Anaconda surveys over the years: in the 2021 survey, one-third of the organisations replied that they have no plan to take any steps towards ensuring fairness and mitigating bias or to address model explainability and an equal percentage that they do know whether there is a plan or not [23]. In 2022, 24% said that they do not employ any measures regarding the same issues; the “not sure” reply was around 15% [24]. In 2023, it was reported that one of the roles that companies hire or planned to hire is AI ethicist [25].

2.3. What Do the Universities Teach?

Since the majority of data scientists have a tertiary degree, it is useful to examine whether they obtain any ELSA education during their studies.

According to the 2018 ACM members survey mentioned above [30], regarding the Computing Competencies for Undergraduate Data Science Curricula, only half (54%) of data science programs required ethics courses, while less than a quarter of the programs actually offered such a course. In a study conducted in Europe in 2020 about the teaching of ethics in computer science or related programs, 36% of the respondents reported that their institutions did not teach ethics. However, the ones that did teach them did so in AI and related areas (64%) and machine learning and related topics (51%) [31].

However, in the Anaconda state of data science data science surveys, only around 20% of the students self-report that ethics in data science/ML is covered in their courses, while a slightly higher amount (22–24%) said that bias in AI/ML/data science is taught frequently in classes or lectures [23,24].

2.4. What the Data Scientists Say about ELSA Challenges

Regarding the attitudes of the data scientists about ELSA issues, we turn to the Anaconda reports; they state that the largest issue to be tackled in AI/ML is the social impacts that stem from bias in data and models, followed by impacts to individual privacy [22–24];

however, when asked about skill gaps, only 15% or the responders identified ethics as a skill lacking from data practitioners at their organisation [23].

The 2023 survey focused specifically on the use of generative AI, and the majority of participants (39%) cited the need for transparency and explainability in AI models, followed by bias and fairness in AI algorithms (27%) and balancing copyright and intellectual property protections with innovation (15%) [25].

2.5. Conclusions

The landscape description regarding ELSA training for data scientists, which we attempted to describe via our review of existing courses and sources describing the profile of the data scientist (according to both industry demands and their own self-perception), led us to the following conclusions:

- Data scientists come both from computer science and statistics disciplines, but a growing number of them also come from other disciplines such as social sciences, finance, and business; however, there is a considerable number that landed in data science following non- or quasi- academic ways.
- Those with an academic education did not necessarily have any ELSA-related courses, either mandatory or elective.
- The industry desires but does not require ELSA knowledge for data science employees.
- A growing number of data scientists recognise the need for awareness and action regarding the ELSA challenges in their work.

3. Curriculum Proposal

Based on our background review to identify the needs an ELSA curriculum for data scientists would address, we composed the proposal summarily presented in this section. Our main concerns were the following:

- Data science projects are multidisciplinary in the following ways: a. the application domain, for example, just examining scholar publications, at least 20 disciplines can be identified [1], b. the implementation of the project requires the cooperation of a variety of experts from many disciplines (computer scientists and engineer, statisticians, visualisation experts, etc.), and c. the impact of these projects make cooperation with other stakeholders necessary, for example legal experts, ethicists, and community representatives (such as patients in healthcare projects).
- ELSA challenges are faced throughout the data science workflow. Some of them run through more than one phase, others not; some may concern all the people involved in the project, others not so much or not to the same degree. In any case, they have to be considered part of the workflow and not as a sometimes optional task.

We tried to address the above concerns in our objectives and vision of the curriculum, leading us to a modular and versatile approach that can be adapted to address different educational needs. We defined the *knowledge unit (KU)* as the elemental component of the curriculum, as a concise piece of knowledge pertaining to a specific ELSA subject. The KUs are situated in a grid created by a horizontal and a vertical partition corresponding to the topics covered and the data science project workflow.

In the following subsections, we state the targets as objectives and vision of the curriculum, and then we present its structure and a brief overview of its constituent parts.

3.1. General Objectives and Vision

Taking into consideration the profile and the ELSA challenges that the data scientists face in their work, we propose an ELSA curriculum, which has three basic objectives. Specifically, data scientists should be able to carry out the following:

1. Recognize ethical, legal, and societal aspects pertaining to their work (awareness);
2. Possess a common language with the relevant domain experts in order to cooperate to find appropriate solutions (communication ability);

3. Incorporate ELSA into the data science workflow. ELSA should not be seen as an impediment or a superfluous artefact but rather as an integral part of the Data Science Project Lifecycle (professional mentality building).

3.2. Curriculum Structure and Overview

As stated above, the curriculum structure has a vertical (modules) and a horizontal (strands) partition, within which the various knowledge units are situated. A KU is the elemental component of the curriculum, defined as a concise piece of knowledge pertaining to a specific ELSA subject.

The horizontal partition is defined by a. the topics identified from the literature review of existing courses [16] and b. the first two objectives (awareness and communication ability); in order to achieve these objectives, data scientists should learn and be able to understand basic concepts belonging to other disciplines (law, ethics, and social sciences), as well as be aware of how ELSA issues could be practically tackled via the use of specific frameworks, standards, and techniques. Thus, we defined three main area topics as follows:

- Ethical and Societal Knowledge Units: Ethical and societal aspects of data science range from incorporating community values into one's work, averting discrimination against individuals or groups, taking into consideration the environmental impact of the data science applications, and assuming responsibility and being accountable for one's decisions and actions.
- Legal Knowledge Units: The objective of these units is to help data scientists cope with legal issues they might be facing in their course of work, mainly data protection and intellectual property issues, as well as basic legal terminology and concepts.
- Technical Renderings Knowledge Units: These units deal with technical renderings of legal or ethical desiderata like privacy, data and algorithmic bias detection and mitigation strategies, incorporation of fairness, effectiveness, and explainability in the evaluation of a data science project, deployment, and monitoring outside experimental/testing environments. They illustrate the way technical solutions can be employed via use cases and do not aim to teach technical skills to data scientists since a. these might be already dealt with in specific courses during their studies or their working experience and b. vary in each application domain that might require specific techniques.

The vertical partition (modules) seeks to achieve the third objective: the seamless integration of ELSA handling in the data science workflow. For this purpose, we employ the CRISP-DM (CRoss-Industry Standard Process for Data Mining), a non-proprietary, documented, and freely available industry-, tool-, and application-neutral model data mining model offered as a best practices guide. CRISP-DM comprises six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment, which altogether provide a road map to follow while planning and carrying out a data mining project [32].

Instead of CRISP-DM, we could have used any other data mining workflow model, such as the Knowledge Discovery in Databases (KDD) [33] and Sample, Explore, Modify, Model, and Assess (SEMMA) [34] models; it is possible to create a correspondence between the workflow phases defined in each model [35], so in principle, any of them could offer us a similar vertical partition. The reasons that we chose CRISP-DM are that a. it is well known and often used in data science projects; b. it has been developed mainly by industry, so it is suited to the needs of the data science practitioners; and c. it is already used to present ethical and legal issues in the different phases of the model and to develop frameworks that ensure the application of ethical standards in the development of data science projects [36–38].

Figure 1 depicts 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.

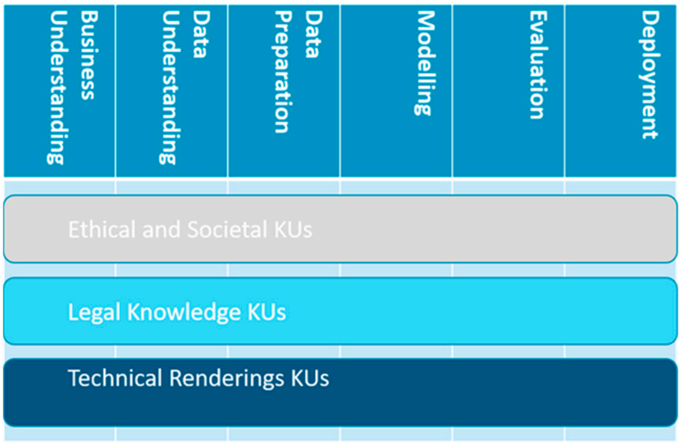

Figure 1. The horizontal–vertical structure of the curriculum. With grey, we denote the ethical/societal strand, with light blue, the legal strand, and with deep blue, the technical renderings of the ELSA challenges. The vertical structure is expressed as modules that correspond to the CRISP-DM phases,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

3.3. ELSA Curriculum KUs

In this section, we will present a more detailed, albeit not exhaustive, description of the KUs. For each module (corresponding to each CRISP-DM phase as described in [32]), we make a summary description of the KUs assigned to it.

The interested reader can find the complete description in Appendix II of [15] (also in Supplementary Materials).

3.3.1. Module I: Business Understanding

This phase concerns understanding the project objectives from a business perspective, converting the problem to a data science problem, and devising the preliminary plan to achieve these objectives.

The corresponding KUs involve stakeholder identification; incorporating community values, such as professional codes of ethics and national and international guidelines, as well as "-by design" guidelines and frameworks for ethics, privacy, data protection, etc.; organisational culture, meaning an organisation's systems, procedures, and practices for guiding and supporting ethical behaviour; basic legal concepts, such as domains of law and specifically of cyberlaw and law stratification (e.g., national vs. international and EU law, and how these relate), as well as concepts regarding accountability measures such as audit, impact assessment, compliance, and risk assurance.

3.3.2. Module II: Data Understanding

This phase concerns data collection, exploration, and quality verification, gaining insights into and identifying possible issues.

This module addresses many issues common in the following module (data preparation), and in specific curriculum implementation, the two modules can be integrated (this approach is also followed in the classification of ethics issues (collectively named data challenges) by [36], which had a major influence in our curriculum structure). However, we create two different modules for the following reasons: a. there are a number of challenges that are to be faced before any pre-processing takes place and that has to be underlined in the curriculum; b. this phase can alter the previous one of business understanding since data understanding can alter the initial problem definition; and c. it might be the case that phases 2 and 3 are not performed by the same people, as data scientists may be provided by already existing datasets and are not directly connected with dataset creation.

That said, depending on the implementation of the curriculum, some of the KUs that deal with common issues can be addressed in either or both modules, as time and purpose constraints vary. For more details, see Section 4 for the implementation strategies.

The KUs in this module deal with data protection and intellectual property issues (basic concepts and issues that have to do with data collection); ways of detecting and mitigating data bias in this phase; the use of synthetic data; ethics dumping; and methods of dataset documentation.

3.3.3. Module III: Data Preparation

This phase aims to construct the final dataset that will be used by the model, a process that includes selection, cleaning, quality checking, integration, and formatting of the collected data.

While the main subjects are same as in the previous model (the KUs bear the same names), we try to differentiate between the relevant issues. That said, depending on the implementation, the respective KUs can be taught independently or not (e.g., in order to avoid repetitions).

The KUs belonging to this module also deal with data protection (here with the application of a specific standard as a framework for a case study) and intellectual property, focusing on the different kinds of training data according to their licences when specifically considering generative AI issues. Also, there are data challenges in pre-processing from annotation, cleaning, and creating synthetic data to choosing features and proxies. Additional challenges include various kinds of bias (representation, measurement, and aggregation bias) or data validity issues: for example, annotation differs depending on whether it is done via crowdsourcing, by experts, or by automatic machine learning systems. Societal issues like employing low-cost solutions for data annotation (e.g., by outsourcing to countries in the Global South) are also addressed here. Dataset documentation is a continuation of the respective KU of the previous module.

3.3.4. Module IV: Modelling

This phase concerns the selection and parameter calibration of models according to the specific problem definition and specific data requirements.

KUs in this module address model bias and possible mitigation techniques, transparency and explainability, environmental impact of model training, intellectual property issues that pertain to the algorithm itself and its products (the use of pretrained models, commercial secrets, property of the trained model, and the outcome of the algorithm), and model documentation frameworks.

3.3.5. Module V: Evaluation

This process includes the review of the model construction and evaluation of whether it achieves the objectives set in the business understanding phase.

The KUs here deal with model evaluation beyond accuracy, i.e., also taking into consideration also factors like fairness, efficiency (in terms of resource allocation), explainability, trustworthiness (robustness outside the experimental settings), and whether the selected features and proxies actually solve the initial problem, although they provide accurate results. Specific focus is given to fairness since the evaluation phase gives us our first tangible result that we can use to assess whether there are bias and discrimination issues. Additionally, we cover here the various kinds of fairness, legal provisions regarding it, as well as human perceptions of fairness and how these impact the trustworthiness of a product and insights from other disciplines (e.g., ethical philosophy) that extend beyond technical solutions. Finally, there is a KU regarding model documentation as a continuation of the previous module KU with the same name.

3.3.6. Module VI: Deployment

This phase includes deployment, monitoring, and maintenance of the system. Even though it is often the customer who carries out the deployment steps, it is important for the customer to understand how they actually use the system and its limitations. As the authors of [36] point out, the data scientist's ethical responsibilities do not end with the completion of a project. The data scientist also has a duty to explain their choices and the implications, using language that non-data scientists, such as managers, can understand.

So, there is a KU regarding the system deployment limitations both with respect to known issues and with what the system is actually able to do or not, and the chosen level of automation and the possible impact that both might have, especially in the case of adverse outcomes. There is also a special KU on visualisation bias and the well-known pitfalls in that area; data scientists are not especially familiar with them as they are usually thought of as UX issues, and this KU aims to help them cooperate with UX designers in order to avoid as much as possible misinterpretation of results. Finally, there is a KU on accountability and processes to ensure it, which is a mirror image of some of the subjects dealt with in Module I: Business understanding, such as auditing frameworks and organisation culture and processes, the extent of personal responsibility, and the case for certification for data scientists in conjunction with professional codes and the creation of a formal profession.

Figure 2 shows a more detailed, albeit not exhaustive, presentation of the curriculum content. Each of these blobs may correspond to more than one knowledge unit; some of the topics addressed in each module can be seen in the figure. To the original three strands, we added a documentation knowledge unit that runs through all modules, emphasising the need for documenting each action taken during the various phases of the project. The way of performing this depends on the phase; we do not advocate for a specific documentation or auditing system.

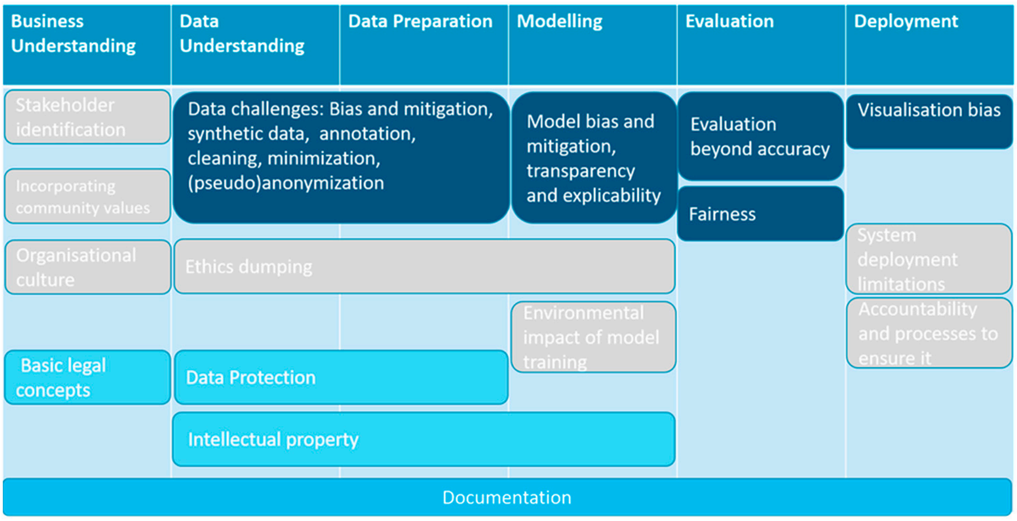
Business Understanding
Data Understanding
Data Preparation
Modelling
Evaluation
Deployment
Stakeholder identification
Incorporating community values
Organisational culture
Basic legal concepts
Data challenges: Bias and mitigation, synthetic data, annotation, cleaning, minimization, (pseudo)anonymization
Ethics dumping
Data Protection
Intellectual property
Model bias and mitigation, transparency and explicability
Environmental impact of model training
Evaluation beyond accuracy
Fairness
Visualisation bias
System deployment limitations
Accountability and processes to ensure it
Documentation

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.

4. Discussion—Implementation Strategies and Limitations

The ELSA curriculum proposal presented so far is a work in progress developing within the framework of a specific research project (FAIR Data Spaces [14]) as one of the project demonstrators. This entails that it has to follow the requirements and ensuing limitations that the project imposes; specifically, the project to connect research institutions with the industry in compliance with the FAIR Principles, i.e., to share data in a findable, accessible, interoperable, and reusable way).

This means that the proposed curriculum is intended for data science practitioners, i.e., people who already have experience in working on data science projects. This is the reason that the background research presented in Section 2 was focused on sketching the data scientist profile and the industry demands regarding ELSA skills.

The target audience level of knowledge and needs is also instrumental regarding the curriculum implementation in deciding program duration and form.

As mentioned in Section 2, data scientists come from a variety of educational and professional backgrounds, fulfilling a variety of roles, ranging from programmers who only implement specific pieces of code to system architects and project managers. Additionally, the application domain may demand a different level and nature of ELSA knowledge, for example, creating a music recommendation algorithm vs. a healthcare system.

The curriculum as presented here is mainly aimed at entry level data scientists, giving them a bird's eye view, so to speak, of some the ELSA challenges they might encounter during their work and may be considered as filling in the knowledge gap of the missing university courses. In this case, the format that might be most effective is an intensive one-week summer school covering all the topics as presented in Section 3, albeit in an introductory manner.

However, there is another way that the curriculum can be implemented: with the objective of fitting the job roles of the participants, the application domain special demands, or both. In that case, the most effective method of implementation might be expanding specific KUs (or contracting others) so that they are aimed at, for example, project managers or developers. An alternative implementation could adapt the content of the KUs to fit specific application domains, for example, NLP, image processing, or healthcare. Regarding the format that would probably be more suited to both cases, the in-depth workshop (or workshops series) addresses specific issues.

Regarding the means of content delivery, i.e., teaching methods or classroom models, in our original review describing the ELSA training landscape, the most commonly referred teaching methods include lectures by faculty staff and guest speakers (either from the academia or the industry) and case- or problem-based studies as the most popular ones, followed by debates, reflection, discussion and role-playing, while the most effective strategies comprise hands-on activities and case studies [16].

The use of case studies will also be employed in the specific proposal. As stated above, the present ELSA curriculum is developed in the framework of a research project (FAIR Data Spaces), a multidisciplinary project that includes as partners legal and ethics experts who cooperate with developers on issues like personal data protection, intellectual property, and informed consent. Specifically in the project are three developed technical demonstrators (biodiversity, data validation and quality assurance, and federated learning-based platform analytics for heath data), which will also be exploited as use cases for the curriculum. An initial suggestion of how these demonstrators can be used can be found in [15] (also found in Supplementary Materials), while a more extensive description will be published with the final curriculum version by the end of 2024.

As stated above, the present ELSA curriculum is a work in progress; evaluating its adequacy is attempted by obtaining feedback by organising workshops and conducting surveys, as well as publicising the relevant project deliverables (referenced in the Supplementary Materials); the comments collected will be used in the second and final version of the curriculum. Naturally, for evaluating the actual efficacy of the curriculum, the best method is to implement a pilot instruction program, followed by a detailed evaluation via participants' and tutors' evaluation sheets or interviews. However, this is not in the scope of the FAIR Data Spaces project.

5. Conclusions

In this paper, we presented a proposed ELSA curriculum for data scientists. To conclude, we underline the following points:

- We tried to draw the profile of the data scientist, starting from the ideal form, as provided in the bibliography, as a multidisciplinary and multi-talented individual and augmenting the profile with the results of empirical studies and surveys as a person coming from a variety of educational and professional backgrounds. This profile reveals that while ELSA issues are considered either by organisations issuing guidelines, the industry, and the data scientists themselves as very important, there is a lack of knowledge about the respective challenges among the practitioners.
- We propose an ELSA curriculum that comprises knowledge units from three domains (ethical, legal, and technical) that belong in modules that themselves correspond to the six phases of the well-known CRISP-DM model.
- The objectives of the curriculum are to raise awareness about ELSA challenges in data science applications; enhance the communication ability of the data scientist by providing a common language with domain experts (for example, legal scholars) that they will have to cooperate in order to successfully tackle the above-mentioned challenges; and finally, to foster a professional mentality that treats ELSA issues as an integral part of the Data Science Project Lifecycle by embedding them into the Data Science Project Lifecycle.
- The implementation of such a curriculum requires a considerable number of resources, both regarding the duration of a training program and its multidisciplinary nature. We propose a flexible implementation strategy that expands, contracts, or conflates KUs in the various modules so as to fit the roles and the specific application domains of the participants. The depth of the subjects treated can vary: for novices, a more holistic but less in-depth program is proposed; for more experienced practitioners, it might be better to focus on specific areas of interest.
- Finally, we published the first version of the curriculum (see Supplementary Materials), and we welcome comments and suggestions from the community. For this purpose, we have already contacted a series of workshops, and currently, we have requested feedback via a specifically designed survey. This will lead to a second version of the curriculum. Naturally, the best way to assess the curriculum will be to actually implement and evaluate it via the experiences of both the instructors and the participants.

Supplementary Materials: The following material can be downloaded from the FAIR Data Spaces Zenodo community (https://zenodo.org/communities/fair-ds (accessed on 26 March 2024)): 'ELSA Training Curriculum for Data Scientists—Version 1.0'. FAIR-DS Project Deliverable. Cologne: UzK, 31 August 2023. https://zenodo.org/record/8318726 (accessed on 26 March 2024); 'ELSA Training for Data Scientists-Describing the Landscape'. FAIR-DS Project Deliverable. Cologne: UzK, 31 December 2021. https://doi.org/10.5281/zenodo.7233569 (accessed on 26 March 2024); presentation slides and video recordings of the ELSA Workshops conducted in 2022 and 2023.

Author Contributions: Conceptualization, M.C. and O.D.B.; investigation, M.C.; writing—original draft preparation, M.C.; writing—review and editing, M.C. and O.D.B.; supervision, O.D.B. All authors have read and agreed to the published version of the manuscript.

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

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article and Supplementary Materials.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Emmert-Streib, F.; Dehmer, M. Defining Data Science by a Data-Driven Quantification of the Community. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 235–251. [CrossRef]
2. Cleveland, W.S. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *Int. Stat. Rev.* **2001**, *69*, 21–26. [CrossRef]
3. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, 1st ed.; Crown: New York, NY, USA, 2016; ISBN 978-0-553-41881-1.
4. Crawford, K. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*; Yale University Press: New Haven, CT, USA, 2021; ISBN 978-0-300-20957-0.
5. 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) (Text with EEA Relevance), 119 OJ L § (2016). Available online: http://data.europa.eu/eli/reg/2016/679/oj/eng (accessed on 26 March 2024).
6. 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 (2021). Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (accessed on 26 March 2024).
7. Jobin, A.; Ienca, M.; Vayena, E. The Global Landscape of AI Ethics Guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
8. 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*; Berkman Klein Center for Internet & Society: Rochester, NY, USA, 2020.
9. AI Ethics Guidelines Global Inventory by AlgorithmWatch. Available online: https://inventory.algorithmwatch.org (accessed on 10 February 2021).
10. *High-Level Expert Group on AI (AI HLEG) Ethics Guidelines for Trustworthy AI*; European Commission: Brussels, Belgium, 2019.
11. Borenstein, J.; Howard, A. Emerging Challenges in AI and the Need for AI Ethics Education. *AI Ethics* **2021**, *1*, 61–65. [CrossRef]
12. Garzcarek, U.; Steuer, D. Approaching Ethical Guidelines for Data Scientists. *arXiv* **2019**, arXiv:1901.04824. [CrossRef]
13. Mittelstadt, B. Principles Alone Cannot Guarantee Ethical AI. *Nat. Mach. Intell.* **2019**, *1*, 501–507. [CrossRef]
14. FAIR Data Spaces | NFDI. Available online: https://www.nfdi.de/fair-data-spaces/ (accessed on 14 February 2024).
15. Christoforaki, M. *ELSA Training Curriculum for Data Scientists—Version 1.0*; UzK: Cologne, Germany, 2023.
16. Christoforaki, M. *ELSA Training for Data Scientists-Describing the Landscape*; UzK: Cologne, Germany, 2021.
17. Davenport, T.H.; Patil, D.J. Data Scientist: The Sexiest Job of the 21st Century. *Harv. Bus. Rev.* **2012**, *90*, 70–76. [PubMed]
18. van der Aalst, W.M.P. Data Scientist: The Engineer of the Future. In *Enterprise Interoperability VI*; Mertins, K., Bénaben, F., Poler, R., Bourrières, J.-P., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 13–26, ISBN 978-3-319-04947-2.
19. Luna-Reyes, L.F. The Search for the Data Scientist: Creating Value from Data. *ACM SIGCAS Comput. Soc.* **2018**, *47*, 12–16. [CrossRef]
20. About Anaconda. Available online: https://www.anaconda.com/about-us (accessed on 14 December 2023).
21. Kaggle: Your Machine Learning and Data Science Community. Available online: https://www.kaggle.com/ (accessed on 14 December 2023).
22. Anaconda | State of Data Science 2020'. Available online: https://www.anaconda.com/resources/whitepapers/state-of-data-science-2020 (accessed on 26 March 2024).
23. Anaconda | State of Data Science 2021'. Available online: https://www.anaconda.com/resources/whitepapers/state-of-data-science-2021 (accessed on 26 March 2024).
24. Anaconda. Anaconda | State of Data Science Report 2022. Available online: https://www.anaconda.com/resources/whitepapers/state-of-data-science-report-2022 (accessed on 26 March 2024).
25. Anaconda. State of Data Science Report 2023. Available online: https://www.anaconda.com/state-of-data-science-report-2023 (accessed on 26 March 2024).
26. Kaggle Kaggle's State of Machine Learning and Data Science 2021. 2021. Available online: https://www.kaggle.com/kaggle-survey-2021 (accessed on 26 March 2024).
27. Zahidi, S.; Ratcheva, V.; Hingel, G.; Brown, S. *The Future of Jobs Report 2020*; World Economic Forum: Geneva, Switzerland, 2020.
28. Di Battista, A.; Grayling, S.; Hasselaar, E. *Future of Jobs Report 2023*; World Economic Forum: Geneva, Switzerland, 2023.
29. 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), Santa Cruz de Tenerife, Spain, 17–20 April 2018; pp. 503–512. [CrossRef]
30. Danyluk, A. *Paul Leidig Computing Competencies for Undergraduate Data Science Curricula-ACM Data Science Task Force*; ACM: New York, NY, USA, 2021.
31. Stavrakakis, I.; Gordon, D.; Tierney, B.; Becevel, A.; Murphy, E.; Dodig-Crnkovic, G.; Dobrin, R.; Schiaffonati, V.; Pereira, C.; Tikhonenko, S.; et al. The Teaching of Computer Ethics on Computer Science and Related Degree Programmes. A European Survey. *Int. J. Ethics Educ.* **2022**, *7*, 101–129. [CrossRef]
32. Shearer, C. The CRISP-DM Model: The New Blueprint for Data Mining. *J. Data Warehouse.* **2000**, *5*, 13–22.

33. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery in Databases. *AI Mag.* **1996**, *17*, 37–54. [CrossRef]
34. SAS Enterprise Miner—SEMMA. SAS Institute Introduction to SEMMA. Available online: https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm (accessed on 14 February 2024).
35. Azevedo, A.; Santos, M.F. KDD, SEMMA and CRISP-DM: A Parallel Overview. In *IADS—DM*; Weghorn, H., Abraham, A.P., Eds.; IADIS: Amsterdam, The Netherlands, 2008; pp. 182–185. Available online: https://www.iadisportal.org/digital-library/kdd-semma-and-crisp-dm-a-parallel-overview (accessed on 26 March 2024).
36. Saltz, J.S.; Dewar, N. Data Science Ethical Considerations: A Systematic Literature Review and Proposed Project Framework. *Ethics Inf. Technol.* **2019**, *21*, 197–208. [CrossRef]
37. Rochel, J.; Evéquoz, F. Getting into the Engine Room: A Blueprint to Investigate the Shadowy Steps of AI Ethics. *AI Soc.* **2021**, *36*, 609–622. [CrossRef]
38. 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*, 2141–2168. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.