Teaching Responsible Machine Learning to Engineers

Anonymous Authors[1]

Abstract

With the increasing application of machine learning models in practice, there is a growing need to incorporate ethical considerations in engineering curricula. In this paper, we reflect upon the development of a course on responsible machine learning for undergraduate engineering students. We found that technical material was relatively easy to grasp when it was directly linked to prior knowledge on machine learning. However, it was non-trivial for engineering students to make a deeper connection between real-world outcomes and ethical considerations such as fairness. Moving forward, we call upon educators to focus on the development of realistic case studies that invite students to interrogate the role of an engineer.

1. Introduction

As machine learning models are increasingly applied in practice, there is a growing interest in the responsible development and use of these models. Although humanities scholars have studied the ethical implications of artificial intelligence for decades, the widespread application of machine learning techniques has opened up new avenues for studying the interaction between the intelligent systems and society. At the same time, major machine learning venues have attracted manuscripts that address the technical challenges of formulating and achieving fairness and explainability in machine learning.

Within and across research communities there is an increased understanding that applying machine learning responsibly is a sociotechnical challenge that should be addressed from multidisciplinary perspectives (e.g., Raji et al., 2021). This sentiment is illustrated in an emergence of new cross-disciplinary conferences (most notably FAccT[1] and AIES[2]) as well as specialized workshops (e.g., Bias and Fairness in AI (Calders et al., 2021)).

It seems imperative that practitioners understand in what ways machine learning models may pose ethical risks and how these risks can be mitigated. Indeed, there is a growing interest to incorporate responsible design in computer science education (Zegura et al., 2020; Raji et al., 2021; Fiesler et al., 2021). However, education on topics of fairness, accountability, confidentiality, and transparency (FACT) geared toward engineers is still in its infancy.

Responsible Machine Learning Education In some programs, ethical considerations are covered as a stand-alone course, emphasizing normative ethical theories. In other programs, ethics may be incorporated as a seminar following a more technical module. While these classes are valuable, they are at risk of divorcing ethical considerations from technical practice (Malazita & Resetar, 2019; Fiesler et al., 2021). As a result, students may have a hard time applying ethical considerations in their daily professional practice as an engineer (Fiesler et al., 2021).

Instead, we believe there is a need to teach responsible machine learning in a way that (1) encourages students to engage with ethical considerations of machine learning systems, (2) is applicable to the daily practice of engineers. With these goals in mind, we have designed a new course, *Responsible Machine Learning* (RML), at *A University*[3], targeted primarily at final-stage undergraduate students majoring in either data science or computer science.

In this paper, we detail the instructional design of the course and reflect upon our experiences. Although RML covered various topics, we will limit our discussion mostly to teaching algorithmic fairness. In the remainder of this paper, we assume the reader is familiar with basic concepts of algorithmic fairness. A recent snapshot of the frontiers of fairness in machine learning research can be found in Chouldechova & Roth (2020).

Lessons Learned After the first course iterations, we found that technical material was relatively easy to grasp for the target audience of our course when it was directly linked to prior knowledge on machine learning. In particular, we found toy examples, demos, and tutorials to be useful tools to foster student understanding.

However, we have also noticed that it is non-trivial for engineering students to make a deeper connection between

[1] https://facctconference.org/
[2] https://www.aies-conference.com/
[3] University name and location are redacted to maintain anonymity.

real-world outcomes and algorithmic fairness. One of the main challenges in teaching RML was to simplify a complex topic to facilitate understanding, without reducing it to a narrow, technical perspective. To this end, realistic and concrete case studies as well as invited lectures have proven to be helpful.

Moving Forward Despite the raising level of public and academic discourse, high-quality educational resources suitable for undergraduate engineering students are scarce. Moving forward, we call upon educators to develop more realistic and concrete case studies, allowing engineering students to connect ethical considerations and technical decision-making in a more meaningful way.

Outline The remainder of this paper is structured as follows. In Section 2, we describe our course design and reflect upon our experiences. In Section 3, we summarize sketch paths for future work.

2. Course Design

Following the principles of constructive alignment (Kandlbinder et al., 2014), our course design consists of three components: learning objectives, learning activities, and assessment. Due to COVID-19 restrictions, the course was taught fully online.

2.1. Learning Objectives

RML is structured around four main themes: Fairness, Accountability, Confidentiality, and Transparency (FACT). Of these themes, fairness and transparency are covered most extensively. The learning objectives of the course were as follows.

At the end of the course, students will be able to:

1. *Evaluate and communicate trade-offs between (socio)technical desiderata of machine learning applications, taking into account diverse stakeholders' perspectives.*

2. *Explain technical and organizational strategies for advancing FACT throughout the machine learning development process.*

3. *Select and implement appropriate strategies for enhancing algorithmic fairness and interpretable/explainable machine learning.*

We would like to highlight a few key aspects of these objectives. First of all, note that learning objective 1 emphasizes *communicating* trade-offs. In practice, even well-intentioned engineers can contribute to harmful technology through implicit design choices throughout the development process. By making trade-offs more explicit, it becomes possible to discuss them with other stakeholders and thereby foster accountability. Second, learning objective 1 emphasizes engaging diverse *stakeholders*, the importance of which as been stressed previously by e.g., Raji et al. (2021). Third, learning objective 2 highlights how different strategies can be applied *throughout* the machine learning development process - not just as an afterthought. And finally, learning objective 3 requires students to *implement* technical evaluation and mitigation strategies, marrying ethical considerations with the daily practice of an engineer.

2.2. Teaching Materials

We have found that high-quality teaching materials geared towards undergraduate engineering students are scarce. Although there exist several graduate-level courses that cover FACT topics in a research seminar format, we consider this format less suitable for undergraduate engineering students. First of all, the target audience of our course may not be able to fully grasp highly technical papers. Moreover, critical position papers typically assume a level of familiarity with the research field that cannot be expected from the target audience of our course.

For RML, we have tried to fill this gap through the development of lectures, lecture notes, and tutorials[4]. Additionally, assigned reading included several chapters of Barocas et al. (2019) (an incomplete work in progress at the time).

2.2.1. Syllabus

We start the course with an introduction to a responsible machine learning process, structured around the CRISP-DM process model (Wirth & Hipp, 2000). In accordance to learning objective 1, our introduction emphasizes the importance of the problem understanding stage. Is this the right problem to solve? Who are the stakeholders of the envisioned system? In particular, we exemplify different types of harm, structured against the moral values they go against (e.g., safety, fairness, transparency, autonomy).

The second module of the course revolves around fairness of machine learning algorithms and the challenges associated with this (learning objectives 2 and 3). We cover several fairness metrics and mitigation algorithms coined by the machine learning community and discuss their limitations.

To facilitate a deeper understanding of the relationship between fairness and technical design choices, we have also developed several Jupyter notebook (Kluyver et al., 2016) tutorials revolving around a case study of Propublica's analysis of COMPAS (Angwin et al., 2016) leveraging several modules of the Python library Fairlearn (Bird et al., 2020). Although these notebooks contain code, their primary purpose is to help students consider the applicability and lim-

[4]Several of our teaching materials can be found on `url-redacted-to-maintain-anonymity` and in (Name, 2021).

itations of fairness metrics and mitigation algorithms in a particular context.

Additionally, invited lectures of researchers and practitioners helped to engage students with contemporary research discussions and showcase challenges data scientists might face in practice.

2.2.2. Fairness as An Optimization Problem

We found that connecting fairness metrics and algorithms with prior knowledge on machine learning helped students to understand technical details. In particular, the usage of toy examples, demos, and code tutorials seemed to increase student understanding.

Many techniques aimed at achieving fairness-by-design in machine learning can be framed in an optimization context (Zafar et al., 2019). Through this lens, the goal is to maintain good predictive performance while satisfying a number of group-level or individual fairness constraints. This can be achieved via several different techniques including fairness-aware representation learning (Zemel et al., 2013; Hu et al., 2020), model induction, model selection, regularization, or post-processing of specific (Kamiran et al., 2010) or any (Hardt et al., 2016) trained models or model outputs. If a student has recently learned about concepts such as cost-sensitive learning, it becomes easier to master these topics. Similarly, prior understanding of trade-offs between predictive performance metrics (e.g., precision and recall or ROC-curve analysis) helps to better understand other trade-offs, such as a fairness-accuracy trade-off or conflicting notions of fairness.

An exception to this is counterfactual fairness (Kusner et al., 2017). As the majority of engineering students have not previously studied the causal inference framework, it proved difficult to teach this notion of fairness in a compact way. However, we found that exemplifying Simpson's paradox can help to understand the relationship between interrelated features and notions of fairness.

2.2.3. Fairness as a Sociotechnical Challenge

As we will expand upon in Section 2.3, it was non-trivial for students to connect technical design choices and real-world outcomes. As such, one of the main challenges in developing teaching materials was to simplify a complex, sociotechnical challenge like fairness into something that can be understood by our course's target audience, without reducing it to a narrow, technical perspective.

For example, historical biases may be encoded in data, which can result in downstream allocation harms. While this is important to understand, reducing unfairness to "bias in, bias out" foregoes many more fundamental questions, such as whether a predictive model should exist at all. Similarly, after covering fairness metrics, an often-heard question is "which fairness metric should I use?". The answer to this question highly depends on the context of an application. However, this is hardly a satisfying answer. To some extent, reducing the complexity of these challenges through general frameworks seems unavoidable, but risks only a surface-level engagement with a context on the student's part.

2.3. Assessment

The assessment of RML consisted of three components, an individual assignment (20%), three quizzes (15%), and a final group project (65%). As most of our findings relate to the individual assignment and group project, we will limit our discussion to these.

2.3.1. Individual Assignment

In the individual assignment, students practiced identifying risks and balancing trade-offs of machine learning systems (learning objective 1). The assignment was inspired by Zegura et al. (2020), who developed two role playing activities in which students need to decide whether a specific artificial intelligence application should be deployed or not. In RML, the individual assignment was in the form of an individual report covering two scenarios, complemented by two group discussions.

The group discussions served to practice communicating trade-offs and exchanging views with peers. Trained primarily as engineers, many of our students were not familiar with instructional formats involving group discussions. To facilitate a fruitful discussion, we provided students with a suggested timing, meeting roles, and general discussion guidelines.

The majority of our students indicated they appreciated the group discussion format, as it allowed them to gain several new insights. This was reflected in their reports: most students were able to identify relevant stakeholders and high-level benefits and risks of the envisioned system. However, students had more difficulty with the precise formulating of risks and mitigation strategies. For example, students would write that the system "should be fair for all patients" or "without bias against minority groups" without exemplifying what "fair" or "without bias" entailed in this specific scenario. Similarly, students sometimes had difficulty connecting (technical) design choices to the identified risks, reflected in ambiguous phrasing of how mitigation strategies might alleviate some of the risks.

2.3.2. Group Project

For the final assessment, we have taken a problem-based learning approach (De Graaf & Kolmos, 2003). In teams

of five, students went through all stages of the machine learning development process (except for deployment) and implemented techniques for enhancing fairness and interpretabilty/explainability (learning objective 3).

The development of a suitable project was highly non-trivial. We believe that developing a realistic prototype is key for students to fully appreciate the challenges of responsible design from the perspective of an engineer. As such, we set out to find a suitable real-world data set accompanied by a realistic scenario. Fairness assessments involve sensitive data, which made it challenging to find an external partner who was wiling to collaborate in the context of undergraduate course work. Additionally, bench-marking data sets that are routinely used in fairness research often lack the necessary context (e.g., the UCI Adult data set) or relate to contested applications of machine learning (e.g., Propublica's COMPAS data set, see Bao et al. (2021) for a detailed account).

Eventually, we settled upon the MIMIC-Extract data set (Wang et al., 2020), a partly preprocessed data set build upon the freely accessible critical care database (Johnson et al., 2016). The associated task was the development of an ICU mortality prediction model that could potentially be used as decision-support tool for physicians. In the assignment, the tool was positioned as a potential alternative to the well-established Sequential Organ Failure Assessment (SOFA) scores.

By design, the assignment was relatively open-ended. Although the scenario hinted towards fairness and transparency, no explicit requirements were given. Instead, students were required to identify requirements through their analysis of the context. To emphasize the importance of the problem formulation, a large proportion of points was awarded to this part of the assignment (learning objective 1). To teach the importance of fostering accountability, students were also required to fill out a data sheet (Gebru et al., 2018) and model card (Mitchell et al., 2019). Finally, students were asked to reflect upon their findings and explicitly reflect upon (ethical) implications of limitations of their developed model.

From the course evaluation survey, it was clear that many students highly appreciated the project. We found that most groups were able to successfully apply various machine learning techniques, including fairness assessment and techniques for enhancing interpretability or explainability. However, similar to the individual assignment, some groups were not able to articulate the relevance of these approaches precisely in the given context. For example, students were able to successfully compute a set of fairness metrics, but did not explain clearly why they believed a specific fairness metric was applicable within the scenario.

3. Moving Forward

With the design of RML, we set out to build a bridge between ethical and technical perspectives, in a way that speaks to engineers. In this paper, we have showcased our approach and reflected upon our experiences. However, much work remains to be done.

3.1. Realistic Case Studies

Although there is an increasing number of examples that showcase how machine learning models can be harmful, it can be difficult for students to connect technical decision-making with ethical implications beyond surface-level observations. As such, we believe that realistic and concrete case studies are crucial to facilitate student learning.

However, it has proven difficult to develop these materials within the context of a single university course. Firstly, the sensitivity of fairness-related data sets makes it challenging to find external collaborators who ware willing to collaborate. Secondly, publicly available data sets often lack the required contextualization, such as a datasheet (Gebru et al., 2018) or a realistic use case. Some of these issues might be alleviated through the use of carefully crafted synthetic data. However, this would still not allow students to engage with stakeholders' perspectives in a meaningful way and instead leave them to rely on their own assumptions about a scenario. One way forward would be to expand a case study not only with a description of the scenario, but also with direct input from (potentially fictional) stakeholders. For example, this could be organized as video-recorded interviews or written testimonials.

3.2. Interrogating the Role of an Engineer

Ethical development of machine learning is a sociotechnical challenge that cannot be solved by engineers alone. In our view, engineering students should not be expected to be well-versed in all these different disciplines. Instead, we believe it is important to show students the limitations of the computer science lens and present concrete approaches to invite other perspectives.

We call on educators to develop more examples of multidisciplinary work that showcase the role of an engineer as well as other actors. For example, Raji et al. (2021) suggest to develop frameworks to cooperate with peers from other disciplines and to engage with affected populations. At engineering universities, organizing team work with other disciplines can be impractical. A different way to reflect the importance of other disciplines in course work would be to give students the opportunity to consult external experts, possibly in the form of auxiliary context that is only provided on demand.

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica, May*, 23(2016):139–159, 2016.

Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., Lum, K., and Venkatasubramanian, S. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks, 2021.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020. URL https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

Calders, T., Ntoutsi, E., Pechenizkiy, M., Rosenhahn, B., and Ruggieri, S. Introduction to the special section on bias and fairness in AI. *SIGKDD Explor.*, 23(1):1–3, 2021. doi: 10.1145/3468507.3468509. URL https://doi.org/10.1145/3468507.3468509.

Chouldechova, A. and Roth, A. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89, April 2020. ISSN 0001-0782. doi: 10.1145/3376898. URL https://doi.org/10.1145/3376898.

De Graaf, E. and Kolmos, A. Characteristics of problem-based learning. *International Journal of Engineering Education*, 19(5):657–662, 2003.

Fiesler, C., Friske, M., Garrett, N., Muzny, F., Smith, J. J., and Zietz, J. *Integrating Ethics into Introductory Programming Classes*, pp. 1027–1033. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450380621. URL https://doi.org/10.1145/3408877.3432510.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems, NIPS 2016*, pp. 3315–3323, 2016.

Hu, T., Iosifidis, V., Liao, W., Zhang, H., Yang, M. Y., Ntoutsi, E., and Rosenhahn, B. FairNN - conjoint learning of fair representations for fair decisions. In *Proceedings of the 23rd International Conference on Discovery Science, DS 2020*, volume 12323 of *LNCS*, pp. 581–595. Springer, 2020. doi: 10.1007/978-3-030-61527-7_38. URL https://doi.org/10.1007/978-3-030-61527-7_38.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Kamiran, F., Calders, T., and Pechenizkiy, M. Discrimination aware decision tree learning. In *Proceedings of the 10th IEEE International Conference on Data Mining, ICDM 2010*, pp. 869–874. IEEE Computer Society, 2010. doi: 10.1109/ICDM.2010.50. URL https://doi.org/10.1109/ICDM.2010.50.

Kandlbinder, P. et al. Constructive alignment in university teaching. *HERDSA News*, 36(3):5, 2014.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., and development team, J. Jupyter notebooks – a publishing format for reproducible computational workflows. In Loizides, F. and Scmidt, B. (eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pp. 87–90. IOS Press, 2016. URL https://eprints.soton.ac.uk/403913/.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Malazita, J. W. and Resetar, K. Infrastructures of abstraction: how computer science education produces anti-political subjects. *Digital Creativity*, 30(4):300–312, 2019. doi: 10.1080/14626268.2019.1682616. URL https://doi.org/10.1080/14626268.2019.1682616.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Raji, I. D., Scheuerman, M. K., and Amironesei, R. You can't sit with us: Exclusionary pedagogy in ai ethics education. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT

'21, pp. 515–525, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445914. URL https://doi.org/10.1145/3442188.3445914.

Wang, S., McDermott, M. B. A., Chauhan, G., Ghassemi, M., Hughes, M. C., and Naumann, T. Mimic-extract. *Proceedings of the ACM Conference on Health, Inference, and Learning*, Apr 2020. doi: 10.1145/3368555.3384469. URL http://dx.doi.org/10.1145/3368555.3384469.

Wirth, R. and Hipp, J. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1. Springer-Verlag London, UK, 2000.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019. URL http://jmlr.org/papers/v20/18-262.html.

Zegura, E., Borenstein, J., Shapiro, B., Meng, A., and Logevall, E. Embedding ethics in cs classes through role play. https://sites.gatech.edu/responsiblecomputerscience/, 2020.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pp. 325–333. PMLR, 2013. URL http://proceedings.mlr.press/v28/zemel13.html.