Norms, Compliance, and Credit: Evaluating EDM in Context

Hayden V. Lawrence & Collin F. Lynch
Department of Computer Science, College of Engineering
North Carolina State University
Raleigh, North Carolina, USA
hvlawren@ncsu.edu; cflynch@ncsu.edu

ABSTRACT

Educational Data Mining has the capacity to tailor the educational experience to each student and to improve student performance. This is not without issues in representing the norms and restraints, ensuring compliance with the norms and restraints, and assigning credit and blame. Work thus far indicates partial solutions in representation, partial definition of restraints, a requirement for multiple means of compliance verification, and imprecise credit and blame assignment.

1. INTRODUCTION

Research in Educational Data Mining has led to the development of robust models for student performance (e.g. [10]), behavior detection (e.g. [13]), as well as content structure (e.g. [26, 7]). These in turn have been applied to a wide range of areas to personalize online content (e.g. [33]), generate assignments (e.g. [14]), provide data-driven tutoring (e.g. [28]), and even guide whole courses (e.g. [1]). While these tools have been successful they have raised a host of new policy issues for students, parents, educators, and educational institutions [4]. These include questions of how to keep educational data private, or what traditional privacy or other ethical controls mean when educational data is transferred across the cloud and shared between individuals, institutions, and service providers [18]. Who is the ultimate authority, or 'owner,' of the data [27]? How can we enforce issues of fairness and transparency on opaque models [15, 8, 32], or ensure that they conform to our normative and regulatory goals [23]? What role do institutions play in mitigating these harms or in governing the spread of data? And how will the rise of data-driven education fundamentally change our conception of the education process or even what liberal education means [12, 9, 31]?

Ultimately these are policy questions which must be addressed by the parties involved. In this paper we present a short survey of prior work that focuses on three overriding questions. First, what *norms* should govern educational data mining? Second, how do we *implement* those norms in rules and in code and then *verify* that our systems comply with them? And third, how do we quantify harms and *assign credit* for successes particularly when multiple actors are involved?

2. NORMS, GOALS, AND RESTRAINTS

All policy is governed by social *norms* which present the overall goals of a policy, "to educate the next generation of scientists and engineers," or set specific limits such as a restriction on segregation of a student or or other rules that: "impair and inhibit his ability to study, to engage in discussions and exchange views with other students, and, in general, to learn his profession"[1]. In order to evaluate the impact of a trained model or other data-driven educational intervention we have to assess it against existing norms. As the field of educational data mining has grown individual models have been evaluated based upon their performance against agreed upon benchmarks of improved near-term outcomes and self-evaluation. However as trained models become more widespread it will be necessary to consider broader norms and more complex tradeoffs of fairness, efficiency, and equality. We therefore have to consider how to represent these normative goals and restraints, and what normative restrictions should be considered in our analysis.

2.1 Representing Norms

According to the U.S. Department of Education, the goal of online education systems is to "achieve greater learning outcomes" while "customizing the learning experience for each student" [11]. While this definition is intuitive it is not one that intelligent agents or existing ML algorithms can understand or validate. As a consequence it does not lend itself either to automatically monitoring of educational or a "bright line rule" rule for data mining that can be implemented in real time. A number of authors have sought to address this exact challenge by developing formal models for implementing and evaluating norms.

Bench-Capon and Mogdil sought to draw a distinction between two classes of formal normative models which they termed shallow and deep. Shallow models are those that require translation by a subject-matter expert, such as in the quantification above[5]. When translating, one often has to consider multiple laws. These laws frequently conflict. One

[1]Decision of Chief Justice Vinson U.S. Supreme Court in the case: McLaurin v. Oklahoma State Regents, 339 U.S. 637 (1950)

potential solution is to prioritize norms from the laws [25]. This approach has multiple opportunities for failure in translation and prioritization, however, it can be useful so long as the limitations of this shallow model are known and others confirm that the model conforms to the policy goals.

More generally, Aldewereld and Vasconcelos proposed that tasks should be tasked with formally quantifying goals in a machine-readable way [2]. This formal representation would identify key factors such as the individual or agent responsible for an outcome or action, the individual taking the action, and the goal of the action. These goals in turn would need to be represented quantitatively (e.g. X% of students must achieve Y score or students must attend all classes). This is inline with the shallow model approach. The alternative, a deep model approach, requires that norms are embedded as intelligent agents which understand goals such as "achieve greater learning outcomes" [5]. While this split framework is more extensible, however, it is also much more resource intensive. And ultimately this representational task remains generally unsolved.

2.2 Key Restraints

Correcting misconceptions is nontrivial, repeating individual lessons takes time, and repeating a school year takes a permanent toll on a child's development and their career prospects. Therefore as autonomous trained agents take a greater role in education we have to consider not just the goals that they are designed to achieve but the restraints that should limit their actions to avoid harm. These harms can arise from giving incorrect or unfair advice, using incorrect resources, or sharing protected information. Some of the strongest opposition to educational data mining arises from a, not invalid, concern that existing systems will not be subject to restraints and will cause real harm due to violations of existing norms, or simple system error. Or that these systems will serve to entrench biased policies or overly rigid structures in ways that are difficult to track [34, 27].

2.2.1 Privacy

As in many countries, U.S. Federal laws on educational data, the "Family Educational Rights and Privacy Act" (FERPA) restricts how the personally identifying information (PII) of students is used [11]. It cannot be used for any purpose other than that which it was shared. However exactly *how* this restriction is interpreted in different contexts varies varies however. Use of student data to advertise goods and services is generally restricted as is linking data to third-party profiles for marketing. Use of data for research generally requires that it be anonymized prior to use. However whether such data can be used to guide commercial data-driven applications, or whether the logfiles and cloud data collected by third parties should be protected is still subject to debate.

In essence this raises important questions both about the control of data and the extent to which involved parties can set limits. If, for example, we view data as property then who owns it? Is it the students who generated it as part of a compulsory exercise, or their legal guardians?; Is it the instructor who guided the classroom activities that are being monitored?; Is it the institution who commissioned it?; or is it the service or infrastructure providers who built the systems that are being used? [27]. These questions matter because data has high value for students, instructors, and institutions who may learn from it or be judged, possibly even fired, because of it. It also has extreme commercial value to service providers who can use it to build advanced systems and to guarantee monopolies. All of these parties have different interests in controlling the spread of data and in verifying its' accuracy, especially when it is used by others. Even anonymized data can raise serious privacy concerns [16]. It is possible to deanonymize data, thus bringing potentials harms to the students, educators, and more. Even if not able to link data to a specific individual, it is possible to create unique profiles which can be used for tracking. Some policy proposals have been made to mitigate potential harms caused by amassing educational data. But this debate is ongoing and gets reset with each new use.

2.2.2 Discrimination and Fairness

Just as school systems, teachers, and volunteers can all discriminate against individuals, so too can algorithms, particularly trained models which can entrench the consequences of prior biases into a seemingly infallible form. It is therefore important for educational data miners to consider the role that biases may play in determining their results. One naïve approach to avoid discrimination would be to mandate transparency for code or decisions. If, for example, code and decisions can be audited then evidence of discrimination can be identified. Such exposure, however, would necessitate the disclosure of trade secrets and it would fail to address the potential of biases within the model or training data [3]. Exposing either one risks violating individual privacy and it still begs the development of a public standard for such biases and the fact that post-hoc auditing still requires that someone face harm.

Kroll et al. proposed one avenue to address this via independently verifiable proofs, which do not require access to source code could verify that some restraint was followed, such as that a teaching agent did not use race or gender as a qualifier. These could in theory be implemented as a form of cryptographic commitment which allows for the system to be audited without exposure of individual or proprietary data via a zero-knowledge proof [23].

There are additional challenges however when dealing with something as broad as discrimination. First, it is not possible or always appropriate to simply ignore discriminatory features such as gender or ethnicity. Indeed in order to evaluate any potential bias or to implement policies such as affirmative action these features are necessary. Second, even if such hot button features are eliminated bias can still be encoded through secondary proxies such as zip code which correlate to race and economic status. One approach to addressing this, proposed by Kroll et al. is "fair affirmative action" which would normalize the data to ensure equal results across well-defined groups [23].

3. COMPLIANCE VERIFICATION

If we can identify the specific normative goals and restrictions which a model must adhere to, and set standards for their performance, it is is still necessary to verify that algorithms comply with them. There are three general approaches that can be taken to compliance verification. We can set strict standards for development and implementation

that guarantee the resulting system will comply (*a-priori verification*). We can observe the behavior of a system as it is used to judge its' actions (*runtime verification*). And we can analyze the results once they are obtained, sometimes over years (*post-hoc verification*). Each approach has strengths and inherent limitations.

3.1 A Priori Compliance Proof

Kamiran, Calders, and Pechnizkiy describe the creation of a decision tree based model that has bounded discrimination [20]. Here potentially-discriminatory traits of the type discussed above are used when training a model. However the outcome of the trained model is strictly limited to ensure that each group is represented fairly. Thus we enforce a mathematical definition of unfairness at the expense of additional training time and a loss of accuracy in representing our underlying data. This approach is generalizable to other quantifiable constraints. Similar discrimination aware approaches have been taken by others in related contexts [6, 35]

Dwork et al proposed a variation of this model that both limits discrimination and minimizes the loss in fairness called "fair affirmative action." Here 'fairness' refers to an individual receiving minimal loss or gain due to normalization. By normalizing data ahead of time through a Lipschitz classifier, which is a mapping of individuals to outcomes such that any 2 individuals who are similar in a task should achieve similar outcomes [15], statistical parity between groups is achieved while the loss to fairness is minimized. While such a classifier is less common within the context of an autonomous educator, a fair affirmative action based classifier could be used in the college admittance process in order to achieve both fairness and equality [15].

A-priori rules of fairness depend upon us having a clear standard for construction and a formal guarantee that our changes will lead to good outcomes. This approach, however, is ill-suited to cases where the standards are not clear in advance or where we do not have the option to implement them. For that it is necessary to consider runtime verification.

3.2 Runtime Verification

The medical field tackles similar issues to those in educational data mining and modeling. In medicine the interactions between factors are often poorly understood, and creating systems that make formal guarantees also means making strong tradeoffs between privacy and accountability. These tradeoffs become more difficult when we must work with anonymized data or data where each individual piece may be anonymous but larger or more rich collections may violate privacy [27]. One approach that has been taken to handle these issues is a "query and response" type system which enforces restrictions on data as it is requested. Here raw even rich datasets may be available for use but as a given model requests more and more data or makes requests that may be combined to violate some privacy threshold their access will be cut off [17]. In mechanisms like this we implement passive external monitors that enforce pre-specified constraints.

Lashey and Beasley, by contrast, describe an active experimental form of auditing in a series of studies conducted to audit hiring practices [24]. Here they evaluated the decisions made by a series of automated review systems by submitting real resumes that matched but for race or other factors. This approach allowed them to sample the systems to assess potential harm. However it required that such harm be clear and that such online experimentation is possible.

Both of these approaches can be used to evaluate models as they work or as they exist in the field. This is particularly important for self-training systems which may be built to predefined standards but which accumulate additional reinforcement over time. In such cases an advance certification is not possible. However both systems also assume that the essential limits, either privacy thresholds or bias, can be detected in the immediate term or occur on an individual basis. When dealing with more complex problems like persistent unfairness or bias, or where decisions take place over weeks or years, these approaches may not be viable, hence the need for post-hoc audits.

3.3 Post-Hoc Audits

Unfortunately, in many cases the only option to detect harm is on a group scale well after the fact. For cases like persistent, albeit subtle unfairness our only option will be to collect cases and outcomes and then to evaluate their results. This evaluation can be done by applying similar a-priori mining methods to those discussed above [6]. Or by establishing set thresholds for violation that would trigger later analysis. Others have proposed that we focus not on the outcomes but the decisionmaking that led to them [21]. Thus in addition to a decision we also require that models, or system developers, offer justifications that can be used to explain the apparent implicit discrimination. Thus if more female students apply to a competitive program at a university while more males apply to a less competitive program, a purely outcome-based measure would present the appearance of gender bias. By accounting for the conditional probabilities of the variables or their likelihoods we can account for the apparent discrimination and thus show a lack of harm.

The problem with post-hoc audits is, as noted above, they can only be conducted well after the fact once actual harm has taken place. For many problems the individual consequences (i.e. lost opportunities or poor educational performance) will already be set and will not be rectified. Additionally many such harms may be subtle or due to a number of factors that makes any correction a challenge.

4. HARMS AND CREDITS

In order to conduct any feasible audits or even to evaluate the impact of a system on an otherwise complex educational environment we must face two other related issues. First, how do we quantify the amount of harm done by a system. How costly is individual bias especially if we are focusing on small-scale decisions like problem assignment? And second, to whom do we give credit for successes or failures, particularly as models are used in concert?

4.1 Harm Quantification

When we identify cases of discrimination or poor decision-making it can often be difficult to quantify the impact. If, for example, an adaptive tutorial system persistently assigns *newarly*-non-optimal problems to a student which just exceeds their zone of proximal development, how much harm does that cause? And what form does it take? If a student simply gets more frustrated but otherwise passes a class, or does so with a lower grade (a B rather than an A) how do we account for that? Can we even detect when an education is poorer but still comparable over the long-term?

One approach to this is to set a-priori bounds on acceptable harm harm ahead of time. Thus we use the same methodology as those for a-priori compliance to set a limit on how much randomness we will tolerate or how far down we will assign a student [20, 15]. This approach however falls back on having a set limit to work with.

Often it is more practical to measure harm after the fact. This can be measured at a gestalt level. It can also be measured in subsets - within the context of college admission, discrimination can be measured by comparing what percentage of African Americans were admitted to a university compared to the overall percentage of admittance [29, 21]. In these cases, however, it is important to consider context when measuring progress towards a goal. A marginalized group doing worse on a test does not necessarily mean the test itself was discriminatory - circumstances and environmental factors can also impact this [22].

4.2 Credit Assignment

The second and far more challenging problem is one of credit assignment. This is a general problem in AI, organizational administration, and in public policy. Education is, by nature, a multi-agent environment in which effects play out over the long-term. It is therefore supremely difficult to assign individual credit or blame for a student's success or failure, much less the performance of an entire class. This problem is only exacerbated by the fact that standards for performance are defined based upon year level learning and long-term knowledge which is affected by every system.

Aldewereld and Vasconcelos propose to address this by a form of micro-assessment [2]. Here we focus on setting standards for each individual task and specifying responsible parties (e.g. parents are solely responsible for attendance). And then we evaluate each involved agent within that box. That approach can work well when we consider isolated tasks (like individual assignments) and where we can conduct the necessary pre-assessments and post-assessments to validate them. This approach is problematic however in that it requires such constant assessments and it ignores the problems of long-term consequences, potential inconsistencies between materials, and the increasingly complex nature of many intelligent agents which have moved from well-contained tutoring systems to more general class monitors and pervasive individual coaches. It also ignores the essential problem of long-term credit assignment and transfer both of which are core features of education.

5. CONCLUSION

Educational data mining, adaptive educational systems, and personalized learning have all become a part of our educational landscape. This rise in adaptive systems and the corresponding rise in data collection raises a number of important policy and technical questions. How do re represent the constraints that these systems must meet? How do we evaluate them? And where do we assign blame when something goes wrong.

While there has been a great deal of prior research in all of these areas far more remains to be done. Proposed models for the representation and evaluation of norms are as yet limited. And there is not always a clear consensus on what the limits should be or whom should decide how to set them. Moreover even if clear standards can be set it is challenging to evaluate against them. For some potential harms (e.g. improper use of protected information) it may be possible to limit harms in advance through restrictions on data usage and advance audits. For others, (e.g. frustration-inducing bad advice) it may be possible to detect the problems as they occur on an individual basis and to bound them if not eliminate them. But for other problems (e.g. embedded bias) there may be no alternative but to check for harms after the fact. In which case mitigation may not be possible. And finally, where we can see success or failure, we face the basic problem of assigning credit or blame.

In this paper we have presented an overview of some of these issues and have highlighted relevant research. Far more remains to be done both to identify the issues that we must address as EDM becomes more mainstream, and to develop robust mechanisms to address it. In this case, as in others, our ability to develop novel adaptive models, has exceeded our ability to really evaluate them in context which poses important challenges for the future.

6. REFERENCES

[1] R. Agrawal, B. Golshan, and E. Papalexakis. Toward data-driven design of educational courses: A feasibility study. *JEDM | Journal of Educational Data Mining*, 8(1):1–21, 2016.

[2] H. Aldewereld, V. Dignum, and W. W. Vasconcelos. Group norms for multi-agent organisations. *TAAS*, 11(2):15:1–15:31, 2016.

[3] M. Ananny and K. Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3):973–989, 2018.

[4] S. Ben-Porath and T. H. B. Shahar. Introduction: Big data and education: ethical and moral challenges. *Theory and Research in Education*, 15(3):243–248, 2017.

[5] T. J. M. Bench-Capon and S. Modgil. Norms and value based reasoning: justifying compliance and violation. *Artif. Intell. Law*, 25(1):29–64, 2017.

[6] B. Berendt and S. Preibusch. Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. *Artif. Intell. Law*, 22(2):175–209, 2014.

[7] Y. Chen, P. Wuillemin, and J. Labat. Discovering prerequisite structure of skills through probabilistic association rules mining. In Santos et al. [30], pages 117–124.

[8] A. Chouldechova and M. G'Sell. Fairer and more

accurate, but for whom? *CoRR*, abs/1707.00046, 2017.

[9] M. Clayton and D. Halliday. Big data and the liberal conception of education. *Theory and Research in Education*, 15(3):290–305, 2017.

[10] J. Cook, C. Lynch, A. Hicks, and B. Mostafavi. Task and timing: Separating procedural and tactical knowledge in student models. In Hu et al. [19].

[11] Protecting student privacy while using online educational services: Requirements and best practices. 2014.

[12] G. Dishon. New data, old tensions: Big data, personalized learning, and the challenges of progressive education. *Theory and Research in Education*, 15(3):272–289, 2017.

[13] S. K. D'Mello, C. Mills, R. Bixler, and N. Bosch. Zone out no more: Mitigating mind wandering during computerized reading. In Hu et al. [19].

[14] Y. Dong and T. Barnes. Evaluation of a template-based puzzle generator for an educational programming game. In B. Magerko and J. P. Rowe, editors, *Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-17), October 5-9, 2017, Snowbird, Little Cottonwood Canyon, Utah, USA.*, pages 172–178. AAAI Press, 2017.

[15] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. *CoRR*, abs/1104.3913, 2011.

[16] C. Dwork and D. K. Mulligan. It's not privacy, and it's not fair. *Stanford Law Review*, 66(35):35–40, 2013.

[17] R. A. Ford and W. N. Price. Privacy and accountability in black-box medicine. *Michigan Telecommunications and Technology Law Review*, 23(1):1–43, 2016.

[18] D. Gasevic, T. Martin, Z. A. Pardos, M. Pechenizkiy, J. C. Stamper, and O. R. Zaïane. Ethics and privacy in EDM. In Santos et al. [30], page 13.

[19] X. Hu, T. Barnes, A. Hershkovitz, and L. Paquette, editors. *Proceedings of the 10th International Conference on Educational Data Mining, EDM 2017, Wuhan, Hubei, China, June 25-28, 2017*. International Educational Data Mining Society (IEDMS), 2017.

[20] F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In G. I. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, editors, *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, pages 869–874. IEEE Computer Society, 2010.

[21] F. Kamiran, I. Zliobaite, and T. Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowl. Inf. Syst.*, 35(3):613–644, 2013.

[22] R. Knopff. On proving discrimination: Statistical methods and unfolding policy logics. *Canadian Public Policy / Analyse De Politiques*, 12(4):573–583, 1986.

[23] J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. Accountable algorithms. *University of Pennsylvania Law Review*, 165(3):633 – 705, February 2017.

[24] J. N. Lashey and R. A. Beasley. Computerizing audit studies. *Journal of Economic Behavior and Organization*, 70(3):508–514, 2009.

[25] R. E. Leenes and F. Lucivero. Laws on robots, laws by robots, laws in robots: Regulating robot behaviour by design. *Law, Innovation, & Technology*, 6(2):194–222, 2014.

[26] R. Liu and K. R. Koedinger. Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning. *JEDM | Journal of Educational Data Mining*, 9(1):25–41, 2017.

[27] C. F. Lynch. Who prophets from educational data mining? new insights and new challenges. *Theory and Research in Education*, 15(3):249–271, 2017.

[28] B. Mostafavi and T. Barnes. Evolution of an intelligent deductive logic tutor using data-driven elements. *I. J. Artificial Intelligence in Education*, 27(1):5–36, 2017.

[29] D. Pedreschi, S. Ruggieri, and F. Turini. Integrating induction and deduction for finding evidence of discrimination. In *The 12th International Conference on Artificial Intelligence and Law, Proceedings of the Conference, June 8-12, 2009, Barcelona, Spain*, pages 157–166. ACM, 2009.

[30] O. C. Santos, J. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, M. C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, and M. C. Desmarais, editors. *Proceedings of the 8th International Conference on Educational Data Mining, EDM 2015, Madrid, Spain, June 26-29, 2015*. International Educational Data Mining Society (IEDMS), 2015.

[31] T. H. B. Shahar. Educational justice and big data. *Theory and Research in Education*, 15(3):306–320, 2017.

[32] M. Skirpan and M. Gorelick. The authority of "fair" in machine learning. *CoRR*, abs/1706.09976, 2017.

[33] K. Spoon, J. Beemer, J. Whitmer, J. Fan, J. Frazee, J. Stronach, A. Bohonak, and R. Levine. Random forests for evaluating pedagogy and informing personalized learning. *JEDM | Journal of Educational Data Mining*, 8(2):20–50, 2016.

[34] T. Z. Zarsky. Law and technology automated prediction: Perception, law, and policy. *Communications of the ACM*, 55(9):33–35, 2012.

[35] R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 325–333. JMLR.org, 2013.