The Right To Be Forgotten and Educational Data Mining: Challenges and Paths Forward

Stephen Hutt
University of Denver
stephen.hutt@du.edu

Sanchari Das
University of Denver
sanchari.das@du.edu

Ryan S. Baker
University of Pennsylvania
rybaker@upenn.edu

ABSTRACT

The General Data Protection Regulation (GDPR) in the European Union contains directions on how user data may be collected, stored, and when it must be deleted. As similar legislation is developed around the globe, there is the potential for repercussions across multiple fields of research, including educational data mining (EDM). Over the past two decades, the EDM community has taken consistent steps to protect learner privacy within our research, whilst pursuing goals that will benefit their learning. However, recent privacy legislation may cause our practices to need to change. The right to be forgotten states that users have the right to request that all their data (including deidentified data generated by them) be removed. In this paper, we discuss the potential challenges of this legislation for EDM research, including impacts on Open Science practices, data modeling, and data sharing. We also consider changes to EDM best practices that may aid compliance with this new legislation.

Keywords

Data Mining, User Modeling, Right to Be Forgotten, Data Privacy, GDPR

1. INTRODUCTION

Data from learners is a critical component of Educational Data Mining (EDM). This data can include demographic information, performance data, and interactions with educational resources such as games, intelligent tutoring systems, and online learning platforms. This data is essential for core goals within EDM research, including contributing to learning theory [5], informing learning interventions [48], creating dynamic and personalized learning technology [30], and informing education policy [3]. The collection and use of learner data raises a number of ethical and legal concerns, including privacy and data security. However, with proper safeguards in place, such data can have significant benefits for both students and educators. By providing valuable insights into student learning, EDM can support the development of more effective educational practices and policies, and ultimately improve student outcomes.

The data that facilitates EDM research can often include personal identifying information (PII) and other protected information. As such, there has been increased attention to privacy protection in recent years. De-identification (removing or obscuring PII from data) has become a standard practice for data sharing. Similarly, researchers have used secure platforms to store and share data that leverage access controls, encryption, and other security measures to safeguard the data. Furthermore, there are also research methods such as Differential Privacy [13, 19], which aims to provide privacy-preserving data analysis by adding noise to the data to mask any information about individuals while preserving the overall trends and patterns. There has been considerable research attention to finding the balance between data privacy and having the data required to drive meaningful insights [32, 51] and creating environments where data can be analyzed in its entirety, without being directly shared [29].

Outside of the EDM community, data privacy concerns are also rising. School districts and public advocates have expressed concerns about the increasing amount of education data becoming available at scale (either for commercial or research use) [42, 58]. Klose et al. [34] note that educational repositories have the potential to contribute to identity theft if hacked, and have shared potential solutions to facilitate the storing of educational data. The Student Data Privacy Consortium, meanwhile, has created a template data agreement between educators and researchers. This template requires that any sharing of a dataset (including deidentified data) must be agreed upon by the local education authority on each occasion [57, 59]. Such measures will undoubtedly protect learners, but are onerous to the point that they will likely limit how much data is actually shared, subsequently limiting the potential for research to benefit students.

More broadly speaking, legislators are also considering the issue of user data and are passing laws that govern how it can be collected, used, and shared. In the United States, the Family Educational Rights and Privacy Act (FERPA) has governed many aspects of educational data since 1974, however, it is more general data privacy laws that may have the most impact on research today. The General Data Protection Regulation (GDPR) in the European Union and the Children's Online Privacy Protection Act (COPPA) in the United States each try to protect users and give them more

control over their data when interacting online. Local governments have also taken steps to legislate around how data might be used, for example, the Colorado Privacy Act and the California Consumer Privacy Act (CCPA) also both contain guidelines on user data, both with regard to storage and sharing. With this trend of increased legal guidance around user data, we must consider how legislation might impact research practice, and how to adjust our research practices accordingly.

One such aspect of legislation that may impact EDM research, and the focus of this paper, is the right to erasure - more commonly called the right to be forgotten. This right, included in GDPR, with variations in other legislation, states that a user may request that their data be removed. Given the high volume of learner data that is central to EDM research, this has the potential to impact our research practices. For example, Such removal of data could impact if a scientific result replicates, or create a ripple effect with those with whom the data has been shared. In the remainder of this paper, we present some of the primary challenges the right to be forgotten may impose upon EDM research so that we may be proactive in addressing and understanding the implications of these laws.

2. BACKGROUND

2.1 Privacy Legislation and the Right to be Forgotten

On May 25, 2018, the European Union (EU) implemented the General Data Protection Regulation (GDPR), a comprehensive data protection law. It seeks to unify data protection laws across the EU and replaced the 1995 EU Data Protection Directive. In addition, it gives EU citizens more control over their personal data [62]. Regardless of whether an organization is based in the EU, it must comply with the GDPR if it processes the personal data of EU citizens [16]. The right to erasure, also known as the "right to be forgotten," (RTBF) is one of the GDPR's most significant provisions, relative to previous legislation. When specific requirements are met, EU citizens have the right to request that their personal data be erased under the RTBF [53, 64]. This may occur when a person withdraws their consent to processing their data, for instance, or when the personal data is no longer required for the purpose for which it was collected.

This legislation gives users more control over how their personal information is shared and formalizes an issue that had already been discussed in the courts. In the 2014 case of Google Spain vs. Agencia Espanola de Protección de Datos (AEPD) and Mario Costeja González, the European Court of Justice determined that a person has the right to ask for the removal of links to personal information from a search engine if the information is unreliable, insufficient, irrelevant, or excessive for the data processing. Furthermore, the court ruled that the data controller (in this case, Google Spain) was required to take reasonable steps to notify third-party controllers (any other organization with which the data was shared) of the individual's request. Due to this decision, Google established a procedure for people to submit RTBF requests known as the "right to be forgotten" form [18]. However, the ruling was not absolute. It could be superseded by other rights and interests, such as the right to information and freedom of expression. With the passing of GDPR, there are still elements of ambiguity of supersedence, for example, if the processing of the data is required to carry out a task in the public interest or the exercise of official authority vested in the controller [8, 36].

Similarly worded legislation has been enacted outside the EU, including in the US, Canada, and Asia. For example, the California Consumer Privacy Act (CCPA) was enacted in the US on January 1, 2020. It grants residents of California certain rights regarding their personal data, including the right to ask for the deletion of personal data held by a company (though with less severe penalties than GDPR) [26, 1]. The Personal Information Protection and Electronic Documents Act (PIPEDA) in Canada governs the private sector's gathering, use, and disclosure of personal information. It does not explicitly address the right to be forgotten. However, according to the Office of the Privacy Commissioner of Canada in 2019, the Act grants individuals the right to access and amend their personal information and the freedom to revoke their consent to its collection, use, or disclosure [50, 37].

Comparably, Singapore's Personal Data Protection Act 2012 (PDPA) governs how businesses collect, use, and disclose individuals' personal information [63, 11]. It grants people the right to withdraw their consent for the collection, use, or disclosure of personal information and access and correct their personal data. Additionally, organizations must delete personal data under section 26 of the act when it is no longer required for the purpose for which it was collected. The common theme across each of these laws is that they provide people more control over their data and the ability to ask for the deletion of data that is no longer relevant or necessary. The GDPR has established a high bar for data protection in the EU. However, with varying levels of legislation across the world, remaining in compliance with each of the varying laws can be challenging (especially if a dataset contains users from multiple locations). This can be especially challenging for researchers striving to share data and provide transparency regarding their scientific methods.

2.2 Replicability Crisis

Replication (in this context) refers to the verification of a scientific study's finding through reproduction, either from the same data, or new data following the same design. The purpose of this process is to better understand the reliability, validity, and merit of a study's findings [15]. A study is deemed reproducible if a research team is able to obtain its original results through the execution of its original method on the original or a comparable dataset [22]. "Reproducibility is a minimum necessary condition for a finding to be believable and informative" [6].

Despite this importance, replication studies remain somewhat rare in education research and in data research more broadly. In a study conducted on 400 previously published works from leading artificial intelligence venues, none of the papers analyzed reported all details necessary to fully replicate their work [24]. In a study conducted on 30 published works on text mining, for example, only one of the studies provided source code to replicate their findings [46]. The re-

port cited lack of access to data, computation capacity, and implementation methods as primary barriers to replication.

The lack of replication leads to a surprisingly large proportion of spurious results being widely reported, as reported by the Open Science Collaboration [47]. In their report, the OSC, an open collaboration of scientists that seeks to improve scientific values and practices, replicated a hundred studies from three top psychology journals. Their study found that 64% of the replications conducted failed to obtain statistically significant results. These findings highlight the importance of replication research and the need to validate published findings. As such, a growing body of research has begun advocating for researchers to take more active steps to facilitate replication through Open Science practices.

2.3 Open Science and Open Data

Recent years have seen increasing movements developing in favor of Open Science and Open Data. The Open Science movement involves a variety of initiatives and values aimed at making scientific research more accessible, more transparent, and more reproducible and replicable. Open Science incorporates a number of different elements, including open (public) access to scientific publications, the use of open-source software, and (of particular relevance to this article) Open Data. Though ideas around Open Science have been around for a considerable time [12], the contemporary Open Science movement arguably dates to the Budapest Open Access Initiative [10], which called for open archives and open access journals. So too, scientific data has been shared publicly for a considerable time [43], accelerating with the advent of the public Internet/World Wide Web [25]. However, a large proportion of scientific data remains inaccessible to other scientific researchers [60], much less the general public.

Within education specifically, the amount of data available openly expanded considerably in the first decade of the 21st century, with repositories such as TalkBank [41] and the Pittsburgh Science of Learning Center DataShop [35]. In recent years, many learning platforms have made their data sets public, and the International Educational Data Mining Society has inaugurated a prize for each year's best publicly available data set. Indeed, this year's conference (2023) includes Open science badges to encourage researchers to share data, and materials, and pre-register their analysis. Increasingly, many funding agencies supporting educational research worldwide have begun to require data management plans, with strong encouragement to make data openly available [44, 17], and a recent Executive Order by the U.S. government mandates open access to scientific publications and open sharing of data starting in 2026 [45]. As such, the already increasing moves within the field towards Open Science and Open Data appear likely to expand considerably in the next few years.

3. RIGHT TO BE FORGOTTEN AND EDM

The right to be forgotten (RTBF) can have a significant impact on the practice of researchers in educational data mining. Under RTBF, all data generated by a learner must be removed from databases upon their request. This can be a difficult and time-consuming task, especially if the data is stored in multiple locations, has been shared with colleagues, or even made publicly available. This can result in a ripple effect where the request to remove the data must be passed on to anyone who received a copy of the data, making it difficult to ensure that the request has been completely satisfied. This could lead to researchers or other data providers stopping data sharing altogether, which would considerably slow research progress and disproportionately impact researchers from communities where funding is more scarce.

Moreover, the data covered by RTBF is extensive. Data that has previously been protected, such as personal identifying information (PII), is covered, but so is any additional data generated by that learner. Interaction data generated as a learner plays an educational game, for example, although in many cases not identifying, would still be covered. Similarly, data from intelligent tutoring software, online learning platforms, or MOOCs would all be covered. Thus, a domino effect of data removal occurs, one that, in collaborative systems, may go beyond an individual learner. There are some that argue that such a broad definition of user data is not required under the legislation, and there that there is ambiguity. To our knowledge, the inclusion of data beyond PII has not yet been tested/challenged in the courts, but such a challenge may well happen in the future. It is also worth noting that despite the lack of testing, many organizations (including the authors' universities and other universities) are acting with this broad definition of user data, which may in time set a precedent outside of the courts.

Placing the right to be forgotten into the context of EDM requires complex planning and execution, given that the removal of a learner's data is not as simple as deidentification. Considering the GDPR legislation specifically, data providers would need to remove all data generated by that learner from databases and shared data sets. In order to mitigate the impact of RTBF on EDM research, it becomes necessary for researchers to keep detailed records of who has access to the data and to plan for the possibility of data removal in the future. By necessity, researchers are also required to keep identifiers for all data so that data can be accurately deleted upon request. This means that datasets that would normally be fully deidentified, must now retain some level of identification, potentially creating additional privacy risks. Some mitigation strategies may include using secure data-sharing platforms that allow for selective data removal and data-sharing agreements that include specific provisions for compliance with RTBF legislation. We do not currently know of any published statistics of how many RTBF requests are being made, however, anecdotally, the third author of this paper holds an administrative leadership role involving handling these requests for their university. Although the university is located in the United States, there have been dozens of requests from EU citizens to be removed, with new batches every month. These requests are then legally required to be processed quickly.

There may also be further impacts of RTBF on research practice. For example, what if the data has been published publicly? What if results have been published, and they can no longer be replicated if the analysis were run again? What if the data is in ongoing use? If a current study can not replicate a past finding, should they compare to the published version of the finding or the finding from the current data set? How can we detect scientific fraud when published

results can no longer be checked? In the remainder of this section, we consider the potential impacts RTBF may have on the field's practices.

3.1 Data Sharing

The right to be forgotten requires that all learner data be removed. If all of the data is stored in one location, this is a somewhat simple (though potentially time-consuming) task. If the data is stored in multiple locations (e.g., multi-site collaborative projects), the task is more challenging and requires slightly more coordination. Should the data have been shared with colleagues outside the immediate collaboration (for purposes of replication or data sharing), it becomes more challenging still, with perhaps the highest challenge being if the data is shared publicly, with no record of who downloaded it.

The right to be forgotten can create a ripple effect, with the request needing to be passed to anyone who received a copy of the data from the original researcher. This effect could result in a significant amount of time required to remove an individual learner's data. This effort increases drastically if the researchers do not have a clear record of who has the data, and it becomes almost impossible if the data was shared publicly. In this case, a researcher could remove the learner's data from the public posting, but not from everyone who had already downloaded a copy, thus not completing their responsibility.

One option to counter the challenge that the right to be forgotten places upon data sharing, is to simply stop sharing data. To stop publishing datasets online or sharing with colleagues. However, this comes with significant disadvantages. Data collection is expensive [14], if data is not shared, data-driven research (such as data mining) will be limited to those that can afford to collect their own data. This will limit much of the research in our field to data owners (i.e., industry and those able to complete primary data collection), or to data sets from countries with less restrictive regulations. Put simply, should data sharing stop, research progress will be slowed, and this slowdown will have a disproportionate impact on researchers from communities where external funding is more sparse (and therefore it is impractical to collect large data sets). Such an equity issue would take the field backwards, and thus we should consider methods that could facilitate data sharing, without creating this particular ripple effect.

3.2 Replicability

Another major ripple effect of the right to be forgotten is in terms of replicability. As noted above, a disappointingly large proportion of research – even machine learning research, where both the data set and code are both available – is not currently replicable [24, 46]. This lack of replicability has several costs. The first and foremost of these is being able to verify if a prior set of analyses was authentic and correctly conducted.

Unfortunately, the right to be forgotten – under certain interpretations – is likely to considerably worsen this problem, and undo the gains of the last several years. If the data set that a past analysis was run on becomes no longer available, it cannot be replicated. Even the removal of one student from a very large data set presents the possibility that a different model will be obtained, or that goodness metrics or statistical results will shift. The field does not currently have methods tailored to determining how much shift could plausibly be expected if one or more students are omitted, and it will be difficult to develop a general framework for predicting shift of this nature, across the broad range of algorithms and models currently used in educational data mining and data science more generally. The field also does not have practices for what to do if – for example – a published finding is no longer obtained within the reduced version of the data set now available. With the right to be forgotten, building on past research will become more difficult and even identifying scientific fraud will be impaired.

Similarly, the right to be forgotten places requirements on data that is "no longer required or relevant" [62]. It is difficult to tell when data is no longer required or relevant, if replication is a future possibility. It is not presently clear if storing data for the purposes of replication will be considered "required" or "relevant" under the legislation. This, in turn, means that further challenges may appear as the practical implications of the legislation (and its interpretation by the courts) become more clear.

3.3 Progressive Science

In addition to replicating previous work, RTBF can present challenges for *building upon* previous work. There is a chance that RTBF requests will result in the deletion of data that is still useful for research [23] - uses that may not be clear at the time of deletion. Similarly, RTBF may limit our ability to compare new work to previous results [4]. For instance, if we cannot replicate prior work, it becomes impossible to tell if a new algorithm is genuinely an improvement upon past work, particularly if a different validation approach is deemed appropriate. Comparative analysis is a crucial technique for assessing the efficacy of different models and pinpointing potential areas for development and future improvement. For instance, a positive recent trend in research on knowledge tracing is the comparison of models across various data sets [20]. This makes it possible for academics and industry professionals to assess the generalizability of their findings, gauge the robustness of new models, and spot data biases or outliers. However, it might be challenging to make these kinds of comparisons and to assess the efficacy of various models if data is removed in response to RTBF requests – two papers could obtain different findings for the same algorithm and supposed same data set.

3.4 Longitudinal Followup

RTBF may also impact the ability to conduct longitudinal studies and monitor student progress. If students exercise their right to be forgotten, comparing and linking data on future outcomes will become more challenging [21]. The goal of longitudinal followup research is often to determine if a curriculum or pedagogy that was effective in the short-term has longer-term benefits for students, particularly students in historically underrepresented groups who are less well-served by current educational systems [52]. Students in historically underrepresented groups are more likely to opt-out of their data being used [40] – in combination with the RTBF, this means that longitudinal research may only be able to demonstrate long-term effectiveness for students who

are already well-served. If an analysis does not explicitly check for consistency of effects across demographic groups, this may lead to an approach being adopted despite (unknown) lower effectiveness for historically underrepresented students.

3.5 Models

One consistent area of EDM research has been the training of statistical and machine-learned models. These models are then integrated into learning environments, dashboards, etc. to provide better learning experiences, and analytics [61]. For example, in [31], models of engagement were trained on data collected from learners, and were later used to create a more adaptive intelligent tutor that responded to student engagement and improved learning [30]. Processes such as these allow the research of the EDM community to directly reach learners and broaden our overall impact.

Currently unclear in legislation is how (and whether) data products are different from the data itself. Consider a machine-learned model from 100 learners' data. That model has embedded in it some representation of the 100 learners. It is likely heavily transformed, and unlikely that the original data could be recreated, but still, the model would be different if only 99 learners' data had been included. The model is a product of the data collected from each of the 100 learners. If a learner enacts their right to be forgotten, must they also be removed from their data's product, the model? Must the model be re-fit?

In large-scale machine learning (such as that conducted by Google), the removal of an individual user likely wouldn't change the model too much. However, given the small Ns often seen in EDM research, the impact could be far greater, and would require increased time on behalf of the research team and place a burden on often limited resources. The difference between data and data product is currently ambiguous in legislation. One interpretation is that an existing model needn't change, but any refinement of the model would need to exclude a learner who had requested to be forgotten. As legislation of this kind becomes more widespread, it is likely that this issue will be considered, and potentially clarified. This clarification may come from legislators, researchers, industry leaders, or the courts. In the meantime, it is important that the EDM community be conscious of this issue, and be involved as data privacy laws are refined. Only by being part of the ongoing discussion surrounding legislation can we ensure that all possible use cases of data are being considered.

4. PATHS FORWARD

The RTBF aims to protect learners and safeguard student privacy, a goal that EDM researchers generally agree with. However, its exact application in EDM has the potential to limit research, and force steps backward in replicability and Open Science Practices. As such, the EDM community should work now to find ways to achieve a balance between research needs (e.g., the need for comparative analysis and data-driven research) and the emerging rights of students to be forgotten.

Given the differences by location, knowing for certain if you are in compliance with privacy legislation can be challenging. We, therefore, advocate for the generation of new best practices in EDM. Such practices could be standardized across the community, and ensure that a researcher is in compliance even with the strictest of RTBF requirements. Striking a balance between research and privacy will not be perfect, but by developing standards as a community, we will have generated a common evaluation point for our research and privacy standards for the learners we work with.

In addressing the challenges described above, we can build on work from colleagues in Healthcare especially [33, 55], where some of these issues have already been tackled. Similarly, we can extend work in our own field that has considered privacy-preserving open science techniques. For example, a recent special issue of the British Journal of Educational Technology reported on technical frameworks for ethical and trustworthy education research [38].

4.1 Privacy-Preserving Live Data Sharing

One possibility for tackling challenges posed to research by RTBF is privacy-preserving Data Sharing. By keeping a live copy of data in a central location, we mitigate a number of the concerns raised in section 3.1. By recording who is authorized to access data, effective logging can be implemented, and downloading or converting can be restricted. An additional benefit of such approaches is they are typically a more accessible way to share data, real-time access to data can be provided without the need for downloading or converting data, which can support those who use assistive technologies.

However, this does not present a perfect solution. Though easier to control the ripple effect of data sharing, implementing the necessary controls to guarantee compliance with these laws may be more challenging. GDPR requires companies handling the data of EU citizens to protect that data, including by implementing privacy by design and by default. Because it can be more challenging to monitor and regulate how data is used in real-time, live data sharing can make it more difficult to comply with these requirements. For instance, it may be challenging to guarantee that authorized users only access data (as opposed to downloading it) or that it is being used for intended purposes. Some of the potential solutions include leveraging cloud services that can satisfy these requirements somewhat easily, as well as the addition of new controls. In such cases, however, a researcher is then relying on a third party to ensure that the solution is compliant. That said, it is not clear if stakeholders (parents, school administrators, etc.) would be in support of the use of third-party data sharing, meaning more exploration of such a solution is needed. Similarly, live data may be susceptible to malicious activity such as hacks - invoking concerns raised in [34]. More research is required to fully comprehend the implications of live data sharing and to determine best practices for overcoming the difficulties presented.

4.2 Privacy Preserving Enclaves

Privacy-preserving enclaves enable the processing and analysis of sensitive data while preserving its integrity and confidentiality [39]. These enclaves isolate a secure environment from the rest of the system using hardware and software based security mechanisms. One such enclave is IntelSGX [54, 56]. To create a secure environment for run-

ning code and storing data, Intel SGX combines hardware and software security features. Even if the rest of the system is compromised, this isolation guarantees that data and computations are shielded from unauthorized access or manipulation [9, 49]. The ability to enable privacy-preserving live data sharing is one of the main advantages of privacy-preserving enclaves. Real-time data processing and analysis are constrained by traditional methods for sharing sensitive data, such as differential privacy or encryption. On the other hand, privacy-preserving enclaves allow sensitive data to be processed and analyzed in a secure setting without compromising the privacy of the people linked to the data.

An EDM-specific example of a privacy-preserving enclave is the MOOC Replication Framework (MORF) which offers a secure environment for the replication and analysis of massive open online course (MOOC) data [29]. Millions of students from all over the world now take part in MOOCs, which have grown in popularity in recent years. However, the data produced by these MOOCs is sensitive, in that students may reveal personal details on discussion forums, which are challenging to perfectly redact at scale [7]. MORF presents a framework for analyzing MOOC data and replicating past analyses without compromising student privacy. MORF allows users to submit analysis code (in any programming language). This code is then run on the MORF database, and the results are provided to the user, without ever having direct access to the data itself. MORF relies upon a variety of security methods implemented within Amazon Web Services, as well as software based protocols that control the output provided to a user (e.g., a user cannot submit a script to extract the data)[2].

Due to privacy concerns, data is frequently kept private in MOOC research, making it challenging to confirm and validate the results of earlier research. MORF provides accessibility for researchers without compromising learner privacy. As such, MORF offers a potential blueprint for privacy preserving data sharing in the future.

These approaches are not without challenges, however. Keeping the underlying hardware and software secure can be a significant challenge. Intel SGX depends on the operating system and hardware security for a secure environment to run code and store data [65]. However, many security flaws in Intel SGX have been found, raising questions about the security of these enclaves. These enclaves' performance is another drawback because privacy-preserving techniques often increase the system's computational and latency overhead, making them less suitable for some use cases. As a result, it's crucial to weigh the trade-offs and ensure that the advantages outweigh the drawbacks. In addition, it can be harder for researchers to work on platforms with the restrictions that privacy-protecting enclaves such as MORF enforce, such as the inability to direct view data or to use unrestricted outputs for debugging. It should also be noted that this approach does not directly address issues of replicability, though it does take steps to prevent the ripple effect.

4.3 Engaging with the Legislative Process

As this legislation evolves and the practicalities are considered and ruled upon in the courts, there will likely be calls for participation from lawmakers, funding organizations, and advocacy groups. Academic research is not something typically well represented by lobbyists [27], thus, we must more actively engage in the process ourselves. This may take many forms, including response to data collection requests (e.g., surveys, interviews, etc.) from legislators, and organizations working on these problems (such as the National Science Foundation). Another form of participation is providing feedback during comment periods for proposed legislation. Engaging with the legislative process provides a better chance that the needs of scientific work, as well as those of the Open Science and Open Data protocols we are encouraging, are considered by lawmakers.

4.4 Collaboration with other disciplines

The EDM community is not the only one facing these challenges. As such, there may be much to learn from how other research areas and industries tackle these challenges. For example, there are already protocols for sharing data in healthcare that satisfy the Health Insurance Portability and Accountability Act (HIPPA), and its privacy rule [28, 55]. Many of these protocols would also facilitate the kind of logging required to satisfy RTBF requests. By taking advantage of existing advancements, we reduce the burden on our research community and avoid 'reinventing the wheel'.

The push for Open Science and Open Data has been a prominent movement across multiple scientific disciplines. The conflicts discussed in this paper, along with the need to find a balance of compliance with legal restrictions and scientific integrity, are not unique to EDM. By working with our research colleagues across disciplines, we can reach more standardized solutions, which would, among other benefits, support standardized requirements regarding Open Science and Open Data in publishing venues, etc. Similarly, other disciplines may benefit from EDM advances in this area, such as MORF [29].

5. CONCLUSIONS

The right to be forgotten, and similar legislative changes on how we store and use data, are likely to have a significant impact on Educational Data Mining. Though we have noted some potential paths forward to adapt to this change, there is not one clear solution. We encourage others in the EDM community to consider the challenges outlined, the potential solutions, and to be proactive, rather than reactive, to these changes. Such proactivity may take multiple forms: it could include designing data-sharing infrastructure, responding to requests for feedback on proposed legislation changes, or joining conversations regarding the interaction of data privacy and research outside our community. A number of advances have been made with challenges similar to these in the healthcare community, and there is much we could potentially learn from other research environments. The EDM community has had a significant impact on learners and education, and has a continued potential to do so. As legislature changes, we must protect that potential, whilst still providing learners with all the protection they can, and should, receive. It is thus our argument that we should develop and adopt best practices now, to be ready for these changes as they are implemented.

6. REFERENCES

[1] J. S. Baik. Data privacy against innovation or against discrimination?: The case of the California Consumer Privacy Act (CCPA). *Telematics and Informatics*, 52, 2020.

[2] R. Baker, S. Hutt, M. Mogessie, and H. Valayaputtar. Research using the mooc replication framework and e-trials. In *2022 IEEE Learning with MOOCS (LWMOOCS)*, pages 131–136. IEEE, 2022.

[3] R. S. Baker. The Current Trade-off Between Privacy and Equity in Educational Technology. In G. Brown III and C. Makridis, editors, *The Economics of Equity in K-12 Education: Necessary Programming, Policy, and Systemic Changes to Improve the Economic Life Chances of American Students*. Rowman & Littlefield., Lanham, MD, In Press.

[4] G. Bansal and F. F.-H. Nah. Internet privacy concerns revisited: Oversight from surveillance and right to be forgotten as new dimensions. *Information & Management*, 59(3):103618, 2022.

[5] G. Biswas, R. S. Baker, and L. Paquette. Data Mining Methods for Assessing Self-Regulated Learning. In *Handbook of Self-Regulation of Learning and Performance*, pages 388–403. Taylor & Francis Group, 2017.

[6] K. Bollen, J. T. Cacioppo, R. M. Kaplan, J. A. Krosnick, J. L. Olds, and H. Dean. Social, behavioral, and economic sciences perspectives on robust and reliable science. *Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*, 1, 2015.

[7] N. Bosch, R. Crues, N. Shaik, and L. Paquette. "Hello,[REDACTED]": Protecting student privacy in analyses of online discussion forums. In *International Conference on Educational Data Mining*. ERIC, 2020.

[8] M. Burri and R. Schär. The reform of the eu data protection framework: outlining key changes and assessing their fitness for a data-driven economy. *Journal of Information Policy*, 6(1):479–511, 2016.

[9] S. Chakrabarti, T. Knauth, D. Kuvaiskii, M. Steiner, and M. Vij. Trusted execution environment with Intel SGX. In *Responsible Genomic Data Sharing*, pages 161–190. Elsevier, 2020.

[10] L. Chan, D. Cuplinskas, M. Eisen, F. Friend, Y. Genova, J.-C. Guédon, M. Hagemann, S. Harnad, R. Johnson, M. L. Kupryte, Rima, I. Rév, M. Segbert, S. de Souza, P. Suber, and J. Velterop. Budapest Open Access Initiative. `https://www.budapestopenaccessinitiative.org/`.

[11] S. Chesterman. After privacy: The rise of facebook, the fall of wikileaks, and singapore's personal data protection act 2012. *Sing. J. Legal Stud.*, page 391, 2012.

[12] D. E. Chubin. Open science and closed science: Tradeoffs in a democracy. *Science, Technology, & Human Values*, 10(2):73–80, 1985.

[13] F. K. Dankar and K. El Emam. Practicing differential privacy in health care: A review. *Trans. Data Privacy*, 6(1):35–67, apr 2013.

[14] G. J. Duncan, N. J. Kirkendall, C. F. Citro, N. R. Council, et al. Data collection costs. In *The National Children's Study 2014: An Assessment*. National Academies Press (US), 2014.

[15] F. Echtler and M. Häußler. Open source, open science, and the replication crisis in HCI. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2018.

[16] EuropeanCommission. 2018 reform of eu data protection rules. `https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf`.

[17] EuropeanCommission. Horizon Europe (HORIZON) Program Guide., 2022.

[18] E. Frantziou. Further developments in the right to be forgotten: The european court of justice's judgment in case c-131/12, google spain, sl, google inc v agencia espanola de proteccion de datos. *Hum. Rts. L. Rev.*, 14:761, 2014.

[19] A. Friedman and A. Schuster. Data mining with differential privacy. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 493–502, 2010.

[20] T. Gervet, K. Koedinger, J. Schneider, T. Mitchell, et al. When is deep learning the best approach to knowledge tracing? *Journal of Educational Data Mining*, 12(3):31–54, 2020.

[21] A. Goldsteen, G. Ezov, R. Shmelkin, M. Moffie, and A. Farkash. Data minimization for gdpr compliance in machine learning models. *AI and Ethics*, 2(3):477–491, 2022.

[22] S. N. Goodman, D. Fanelli, and J. P. Ioannidis. What does research reproducibility mean? *Science translational medicine*, 8(341):341ps12–341ps12, 2016.

[23] E. Gratton and J. Polonetsky. Droit a l'oubli: Canadian perspective on the global right to be forgotten debate. *Colo. Tech. LJ*, 15:337, 2016.

[24] O. E. Gundersen and S. Kjensmo. State of the art: Reproducibility in artificial intelligence. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[25] S. S. Hale, M. M. Hughes, J. F. Paul, R. S. McAskill, S. A. Rego, D. R. Bender, N. J. Dodge, T. L. Richter, and J. L. Copeland. Managing scientific data: the emap approach. *Environmental Monitoring and Assessment*, 51(1):429–440, 1998.

[26] E. L. Harding, J. J. Vanto, R. Clark, L. Hannah Ji, and S. C. Ainsworth. Understanding the scope and impact of the california consumer privacy act of 2018. *Journal of Data Protection & Privacy*, 2(3):234–253, 2019.

[27] P. Harris and C. McGrath. Political marketing and lobbying: A neglected perspective and research agenda. *Journal of Political Marketing*, 11(1-2):75–94, 2012.

[28] T. Hulsen. Sharing is caring—data sharing initiatives in healthcare. *International Journal of Environmental Research and Public Health*, 17(9), 2020.

[29] S. Hutt, R. S. Baker, M. M. Ashenafi, J. M. Andres-Bray, and C. Brooks. Controlled outputs, full data: A privacy-protecting infrastructure for MOOC data. *British Journal of Educational Technology*, 53(4):756–775, 2022.

[30] S. Hutt, K. Krasich, J. R. Brockmole, and

S. K. D'Mello. Breaking out of the lab: Mitigating mind wandering with gaze-based attention-aware technology in classrooms. In *CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[31] S. Hutt, C. Mills, S. White, P. J. Donnelly, and S. K. D'Mello. The Eyes Have It: Gaze-based Detection of Mind Wandering during Learning with an Intelligent Tutoring System. In T. Barnes, M. Chi, and M. Feng, editors, *The 9th International Conference on Educational Data Mining. International Educational Data Mining Society.*, pages 86–93, 2016.

[32] M. Ivanova, G. Grosseck, and C. Holotescu. Researching data privacy models in elearning. In *2015 International Conference on Information Technology Based Higher Education and Training (ITHET)*, pages 1–6, 2015.

[33] M. Jayabalan and M. E. Rana. Anonymizing healthcare records: A study of privacy preserving data publishing techniques. *Advanced Science Letters*, 24(3):1694–1697, 2018.

[34] M. Klose, V. Desai, Y. Song, and E. Gehringer. Edm and privacy: Ethics and legalities of data collection, usage, and storage. In *International Conference on Educational Data Mining*. ERIC, 2020.

[35] K. R. Koedinger, R. S. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper. A data repository for the edm community: The PSLC DataShop. *Handbook of Educational Data Mining*, 43:43–56, 2010.

[36] C. Kuner. The european commission's proposed data protection regulation: A copernican revolution in european data protection law. *Bloomberg BNA Privacy and Security Law Report (2012) February*, 6(2012):1–15, 2012.

[37] D. Lackey and N. Beaton. The current state of data protection and privacy compliance in canada and the usa. *Applied Marketing Analytics*, 4(4):355–359, 2019.

[38] D. Ladjal, S. Joksimović, T. Rakotoarivelo, and C. Zhan. Technological frameworks on ethical and trustworthy learning analytics. *British Journal of Educational Technology*, 53(4):733–736, 2022.

[39] T. Lee, Z. Lin, S. Pushp, C. Li, Y. Liu, Y. Lee, F. Xu, C. Xu, L. Zhang, and J. Song. Occlumency: Privacy-preserving remote deep-learning inference using sgx. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–17, 2019.

[40] W. Li, K. Sun, F. Schaub, and C. Brooks. Disparities in students' propensity to consent to learning analytics. *International Journal of Artificial Intelligence in Education*, 32(3):564–608, 2022.

[41] B. MacWhinney, S. Bird, C. Cieri, and C. Martell. Talkbank: Building an open unified multimodal database of communicative interaction. In *The Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA).

[42] National Association of Secondary School Principals (NASSP). Student Data Privacy, Feb. 2017.

[43] National Research Council et al. Committee on scientific accomplishments of earth observations from space. 2008. earth observations from space: The first 50 years of scientific achievements.

[44] National Science Foundation. Chapter XI.D.4. In *Proposal & Award Policies & Procedures Guide (PAPPG)*. NSF, 2021.

[45] A. Nelson et al. Memorandum for the heads of executive departments and agencies: Ensuring free, immediate, and equitable access to federally funded research. *Repository and Open Science Access Portal (ROSA)*, 2022.

[46] B. K. Olorisade, P. Brereton, and P. András. Reproducibility in machine learning-based studies: An example of text mining. In *ICML 2017 RML Workshop Reproducibility in Machine Learning*, 2017.

[47] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[48] V. Owen, M. H. Roy, K. P. Thai, V. Burnett, D. Jacobs, E. Keylor, and R. S. Baker. Detecting wheel-spinning and productive persistence in educational games. In *EDM 2019 - The 12th International Conference on Educational Data Mining*, 2019.

[49] R. Pires, M. Pasin, P. Felber, and C. Fetzer. Secure content-based routing using intel software guard extensions. In *The 17th International Middleware Conference*, pages 1–10, 2016.

[50] J. Pushka. The applicability of the personal information protection and electronic documents act to de-indexing internet search engine results. *Asper Rev. Int'l Bus. & Trade L.*, 19:175, 2019.

[51] J. R. Reidenberg and F. Schaub. Achieving big data privacy in education. *Theory and Research in Education*, 16(3):263–279, 2018.

[52] T. R. Sass, R. W. Zimmer, B. P. Gill, and T. K. Booker. Charter high schools' effects on long-term attainment and earnings. *Journal of Policy Analysis and Management*, 35(3):683–706, 2016.

[53] B. Sealey et al. Has the 2016 general data protection regulation really given consumers more control over their personal data? *LJMU Student Law Journal*, 1:17–41, 2020.

[54] J. Seo, B. Lee, S. M. Kim, M.-W. Shih, I. Shin, D. Han, and T. Kim. Sgx-shield: Enabling address space layout randomization for sgx programs. In *NDSS*, 2017.

[55] B. Shen, J. Guo, and Y. Yang. Medchain: Efficient healthcare data sharing via blockchain. *Applied Sciences*, 9(6), 2019.

[56] M.-W. Shih, S. Lee, T. Kim, and M. Peinado. T-sgx: Eradicating controlled-channel attacks against enclave programs. In *NDSS*, 2017.

[57] Student Data Privacy Consortium. Standard Student Data Privacy Agreement (NDPA Standard Version 1.0) Version 1r7. Technical report, Access of Learning Community, 2021.

[58] Student Privacy Compass. Student Privacy Primer. `https://studentprivacycompass.org/resource/student-privacy-primer/`, 2021.

[59] F. Tazi, S. Shrestha, D. Norton, K. Walsh, and S. Das.

Parents, educators, & caregivers cybersecurity & privacy concerns for remote learning during covid-19. In *CHI Greece 2021: 1st International Conference of the ACM Greek SIGCHI Chapter*, pages 1–5, 2021.

[60] L. Tedersoo, R. Küngas, E. Oras, K. Köster, H. Eenmaa, Ä. Leijen, M. Pedaste, M. Raju, A. Astapova, H. Lukner, et al. Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, 8(1):1–11, 2021.

[61] K. Verbert, E. Duval, J. Klerkx, S. Govaerts, and J. L. Santos. Learning analytics dashboard applications. *American Behavioral Scientist*, 57(10):1500–1509, 2013.

[62] P. Voigt and A. v. d. Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Publishing Company, Incorporated, 1st edition, 2017.

[63] B. Wong YongQuan. Data privacy law in singapore: The personal data protection act 2012. *International Data Privacy Law*, 2017.

[64] R. Zaman and M. Hassani. Process mining meets gdpr compliance: the right to be forgotten as a use case. In *International Conference on Process Mining - Doctoral Consortium, ICPM-DC*, 2019.

[65] C. Zhao, D. Saifuding, H. Tian, Y. Zhang, and C. Xing. On the performance of intel sgx. In *13th Web Information Systems and Applications Conference (WISA)*, pages 184–187. IEEE, 2016.