The unbearable lightness of consent: mapping MOOC providers' response to consent

Khalil, Mohammad; Prinsloo, Paul; Slade, Sharon

Publication date
2018

Document Version
Final published version

Published in
the Fifth Annual ACM Conference on Learning at Scale

Citation (APA)
Khalil, M., Prinsloo, P., & Slade, S. (2018). The unbearable lightness of consent: mapping MOOC providers' response to consent. In the Fifth Annual ACM Conference on Learning at Scale [61]

The unbearable lightness of consent: Mapping MOOC providers' response to consent

Mohammad Khalil
Delft University of Technology
Delft, Netherlands
m.f.d.khalil@tudelft.nl

Paul Prinsloo
University of South Africa
Pretoria, South Africa
prinsp@unisa.ac.za

Sharon Slade
The Open University, UK
Milton Keynes, UK
sharon.slade@open.ac.uk

ABSTRACT

While many strategies for protecting personal privacy have relied on regulatory frameworks, consent and anonymizing data, such approaches are not always effective. Frameworks and Terms and Conditions often lag user behaviour and advances in technology and software; consent can be provisional and fragile; and the anonymization of data may impede personalized learning. This paper reports on a dialogical multi-case study methodology of four Massive Open Online Course (MOOC) providers from different geopolitical and regulatory contexts. It explores how the providers (1) define 'personal data' and whether they acknowledge a category of 'special' or 'sensitive' data; (2) address the issue and scope of student consent (and define that scope); and (3) use student data in order to inform pedagogy and/or adapt the learning experience to personalise the context or to increase student retention and success rates.

This study found that large amounts of personal data continue to be collected for purposes seemingly unrelated to the delivery and support of courses. The capacity for users to withdraw or withhold consent for the collection of certain categories of data such as sensitive personal data remains severely constrained. This paper proposes that user consent at the time of registration should be reconsidered, and that there is a particular need for consent when sensitive personal data are used to personalize learning, or for purposes outside the original intention of obtaining consent.

Author Keywords

consent; sensitive data; Massive Open Online Courses (MOOCs); learning analytics; privacy policy; terms and conditions.

L@S 2018, June 26–28, 2018, London, United Kingdom
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5886-6/18/06…$15.00
https://doi.org/10.1145/3231644.3231659

ACM Classification Keywords

• **Applied Computing** → Education; • **Social and professional topics** → Computing / technology policy

INTRODUCTION

The provision of consent for the collection, analysis and use of personal data is an issue of little significance for many users and providers of online services despite increasing concerns. Research shows that few users read Terms and Conditions before accepting or opting in to online platforms [3, 35]. In many cases agreeing to Terms and Conditions is a prerequisite for using the service and consent is given lightly. In stark contrast to this apparent 'lightness' are concerns around the increasing pervasiveness of surveillance and algorithmic decision-making [13, 38]; the commercialization of personal data [7] digital promiscuity [42], the reality of "digital serfdom" [23], and the implications of downstream use of personal data by an array of data brokers and role players [1, 5].

It is important to set discourses around the collection, analysis and use of student data in the context of broader discourses on big data and the "data revolution" [41], and the growth in the volume, granularity, velocity and veracity of student data. Of the central concerns in the collection and analysis of data three prominent issues are personal data, issues of consent and use. While oversight of the ethical implications underpinning the collection, analysis and use of student data in higher education remains largely unresolved [21], it is an even bigger issue in the context of informal or post-formal learning contexts such as Massive Open Online Courses (MOOCs). MOOC providers collect, analyse and use student data across platforms, services and continents, and the ethical implications underlying these data assemblages increasingly require scrutiny [1].

While there are processes overseeing the use of participant data in *formal* research environments, it is less evident that such processes apply to the collection, analysis and use of user data in formal higher education where user consent remains a pertinent issue [21, 32, 36].

In the *informal* online learning context, such as MOOCs, it is timely to investigate ways that providers in different contexts approach the issue of student data, and more specifically, sensitive student data, their use of such data and how data are used to shape the student experience [20].

While the issue of user consent in the context of MOOCs has been the focus of existing research (e.g., [35, 48]), there have been important changes since, such as (1) the impact of different legal and regulatory frameworks on student data in inter-continental education delivery; (2) downstream use of collected data for purposes not foreseen at the time of its collection [1]; (3) the "black box" of algorithmic decision-making and machine learning and the limitations/lack of oversight [12]; and (4) developments in the regulatory environments such as the General Data Protection Regulation (GDPR) coming into effect in the European Union in 2018.

This study reviews the practices of 2 MOOC providers from the US and 2 providers from Europe, and asks:

Question 1:
How is 'personal data' defined and is the category of 'sensitive data' acknowledged?

Question 2:
Is student consent addressed, and what is its scope?

Question 3:
Is student data used to change or adapt the learning experience to personalise the context and/or increase student retention and success rates?

It is accepted that educational providers have a fiduciary duty to collect, analyse and use student data in order to deliver effective and sufficiently supported learning experiences [36]. This research investigates the notion, practice and timing of obtaining consent for the context-appropriate use of sensitive, personal data, and questions the appropriateness of large scale collection for less relevant purposes.

It is possible that, at the time of the initial consent, neither student nor provider can foresee the exact scope of data to be collected or all potential uses of that data. However, the granting of initial consent may be seen as providing sufficient rationale for the ongoing collection, analysis and use of data even if the nature of the data has changed since consent was provided. There is an argument then that current frameworks do not sufficiently address the impact of context of the sensitivity of personal data, the scope of consent and the resultant use of the data to inform interventions.

As such, this paper contributes an informed rationale for re-considering the practice, the scope and the timing of consent in specifically educational settings.

DATA, CONSENT AND INTERVENTION

Data has become highly commodified and is perhaps the most highly valued asset of the 21st century [45]. It is broadly accepted that online providers collect and use personal data, ranging from directed, human observation and data collection to automated, increasingly algorithmic-driven processes with(out) user knowledge and consent.

In the context of the "borderless web" [27] and the pervasiveness of platforms like Facebook and Google, the collection, analysis and use of user data have become hotly contested.

The scope of literature and research regarding the implications and scope of collection, analysis and use of data are wide, ranging from, inter alia, surveillance and privacy studies, sociological and philosophical approaches as well as legal frameworks [6, 13, 26, 41]. In this section a brief overview of each of the three central theoretical constructs used in this research is provided, namely 'personal data', 'consent' and 'intervention.' In the context of the collection, analysis and application of user data in education, the frameworks and conceptual thinking set out by Jisc [31, 32] are of specific importance. An analysis of aspects of the Jisc frameworks [31, 32] is set against an alternative consent framework proposed by Cormack [1]. While much of the research on consent critiques existing processes, definitions and legalities, there are few alternative frameworks to provoke a re-think of consent. The selection of Cormack's framework [1] aims to illustrate *one* possibility for re-considering consent, its scope and the processes surrounding obtaining consent.

Within formal (higher) education, there is evidence of a wide acceptance of the inherent potential and examples of a more nuanced approach to the collection, analysis and use of student data [35]. Learning analytics makes it possible to not only offer individualised and personalised support, but to increasingly adapt the learning experience to these identified needs, potential and risks.

Given the nature of international online education provision whereby the institution and its students may be in different geopolitical/legal locations, the collection, analysis and use of student data requires careful scrutiny. Legislative frameworks such as the EU GDPR (http://www.eugdpr.org/) require a (re)consideration of definitions of personal data, the scope and purpose of its collection, and the scope and implications of consent.

Personal Data

The GDPR defines 'personal data' as: "any information relating to an identified or identifiable natural person ('data subject') ... such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person (Article 4)".

Of specific interest here is the distinction between *sensitive* and non-*sensitive* data on which Jisc's framework [31] is based. The GDPR does not formally define 'sensitive data', but states that "Personal data which are, by their nature, particularly sensitive in relation to fundamental rights and freedoms merit specific protection as the context of their processing could create significant risks to the fundamental rights and freedoms" (Article 1:51). It acknowledges then that in some contexts an individual's fundamental rights and freedoms may be compromised. As example, the GDPR mentions ethnic origin and biometric data.

The GDPR allows for 'sensitive' data to be collected, analysed and used on condition of explicit consent or when within the ambit of the legitimate activities of the collecting institution. This notion of 'sensitive' data can be ambiguous and context-dependent. For example, the International Working Group on Data Protection in Telecommunications [20] states that we should not underestimate the sensitivity of digitized student data:

"Personal data about learning behavior may be viewed as particularly sensitive, as these data contain information about the interests and abilities of students, how well they memorize facts, how quickly they can solve exercises of all kinds, and how willing they are to learn something new. Combined with data analytics, they might also be used to predict professional future and career opportunities (p. 2)".

Sclater [32] suggests that attributes such as a person's religion, ethnicity, health, trade union membership, etc., may be irrelevant within the scope of learning analytics. However, he goes on to state that use of characteristics such as race, location or socioeconomic status may be justified should evidence suggest an associated academic risk. In such cases, "The student should also be told exactly what the data will be used for: they have a right not to provide this data, or to have it excluded from a particular type of processing"[32]. The International Working Group on Data Protection in Telecommunications [20] also states that "Educational institutions and providers of e-learning platforms should collect only as much pupil or student data as they need to complete specified purposes" (p. 6).

Consent

Solove [6] suggests that "consent is far more nuanced, and privacy law needs a new approach that accounts for the nuances without getting too complex to be workable" (p. 1901). In the context of this paper, a more nuanced understanding of consent will reflect our understanding of 'data', and in particular, 'personal data'.

The importance of 'context' pertaining to privacy and consent has been highlighted by scholars such as Nissenbaum [17, 18] and Zimmer [28]. Nissenbaum [17] introduced the notion of "contextual integrity" (p. 582) to illustrate and foreground the fact that the conditions informing the sharing of information in one context for a particular purpose fundamentally change when the same information is shared in a different context and possibly for a different purpose.

Much of the discourse surrounding the role of evidence-based decision making in education and the ethical dimensions of research originates in the field of medicine and a biomedical model of ethics as enshrined in the Belmont principles [15, 44]. In more recent contexts, Luger, Moran and Rodden [10] ask "are current notions of consent relevant in the emerging class of ubiquitous systems?" (p. 529). In ubiquitous networks, it ought not to be assumed that users who provide data (as in active sharing or consent), or whose data are automatically collected by algorithms always know at the point of consent or use of the digital service, inter alia: (1) the identity of data collectors; (2) the scope and purpose of collection; (3) the downstream uses of that data; (4) and the potential for third party use [41]. This raises the question of how users provide meaningful blanket consent when there are so many uncertainties at the moment of consent. And while providers may assume the active consent of users as a rational decision in exchange for services, the matter is indeed more complex. There is ample research which challenges assumptions underpinning user consent, such as the rationality of consent in the context of the exchange of information, issues of user control of data and trust [35].

It falls outside the scope of this paper to provide a history of consent (see [37, 44]). Suffice to flag concerns that the "data revolution" [45] presents "unprecedented challenges to how we currently elicit, secure, and sustain user consent" [11, p. 613]. In the context of higher education, Willis, Slade and Prinsloo [21] point out that although it is accepted that learning analytics has ethical implications, there is little oversight and few formal processes relating to its actual application. So how do institutions deal with the issue of consent in the context of learning analytics?

The GDPR defines 'consent' as "freely given, specific, informed and unambiguous indication of the data subject's wishes ... to the processing of personal data relating to him or her" (Article 4). Under the provisions of the GDPR, "pre-ticked boxes... would not be sufficient. A record must also be kept of how and when the consent was provided. In addition, students will have the right to withdraw their consent at any time." In situations where consent serves as condition for using a specific service (see [35]), the GDPR "strongly disapproves" [32]. The providing/processing institution should make clear the purposes and scope of consent at the time of its provision. This means that the scope and purposes for the collection, analysis and use of data are theoretically 'frozen'. "New types of analysis or intervention cannot be added if they were not envisaged at the time consent was obtained" [32]. Given the dynamics and flux in pedagogy and learning, this could mean that the original scope of consent is set so broadly that it would include everything, or even that newer developments requiring different approaches are constrained.

Intervention/purpose

If there is no purpose for collecting data, why would we? In the context of this paper, it is relevant to note that though data are often collected for improving user experience or, in the context of higher education, to improve student retention and success, such operational activity is not seen as 'formal' research. As such the accepted conventions to ensure consent and define the purposes of the collection and use of data at the point of consent do not apply [21]. Uses of collected data in formal research environments are clearly defined, but this is not always the case for digital service provision. In big data environments the purpose of collection and analysis of data is not typically directed by a hypothesis [1]. Digital service providers may collect data not knowing what they will find, or how it might impact on the experience of users who have given consent for its collection and use. From a user's perspective it is crucial to understand consent in terms of 'use' or 'intervention'.

A second issue is the widely accepted use of data by online platforms as a means of improving the general user experience. Increasingly, data are used to shape and personalise specific users' experiences, and limit or expand options, based on specific criteria [4, 22, 34]. And more recently, decisions on what data to collect, how to combine that data with other data sources and how to intervene in the user experience are determined as part of algorithmic decision making [30] with those algorithms running autonomously, and continuously learning in the process [33].

So, when users provide consent to have their data (often undefined in scope) collected and used for undefined purposes and with no clear idea of how this might impact on their user experience, consent sets in motion a process that very few users understand, and even fewer platform providers explain [13].

In a higher education context, the accepted purpose of learning analytics is to improve the chances of student success [8], and despite concerns [40] there is evidence that learning analytics are used to inform pedagogy, allocate resources and inform institutional strategy [2].

In the context of informal learning and MOOCs, there is ample evidence that student data are used to measure and report on engagement, retention and success, but less published research on how *uses* of this data inform pedagogical approaches, change student behaviours or adapt/personalise the learning experience. For instance, Khalil, Taraghi and Ebner [30] raised six concerns related to the employing of learning analytics in MOOCs: informed consent, security, storage of students' data and log files, privacy, ownership, and transparency.

METHODOLOGY

This research is a mixed method interpretative or hermeneutic study [51] using as research strategy the dialogical multi-case study methodology. A dialogical case study attempts to see how a particular theory or a set of theories apply in a particular case [39, 43].

The study furthermore falls in the broader category of instrumental case studies [14], aiming at gaining insight into the ways that four selected MOOC providers define personal data, approach the issue of user consent and use student data to improve students' learning. In following Yin [43], research questions/propositions were identified, cases and units of analysis were selected, and analyses of data linked to the research questions/propositions. In order to achieve a succinct but in-depth understanding of these aspects, a directed content analysis with quantitative and qualitative elements is used [46]. Based on the research questions and literature review, three categories of 'personal data', consent, and data use (intervention) were identified.

Validity, reliability and trustworthiness were addressed by transparency regarding the process including the selection of analytical constructs from the literature review, construct formulation and coding, member checking of the codes, constructs and analyses [46]. Despite the general acceptance of the case study as an accepted and distinct form of empirical inquiry, there are generally two concerns, namely a perception of the lack of rigour and the claim that case studies "provide little basis for scientific generalization" [43, p.15]. The different analyses were imported into a Google document. Each author then commented and verified interpretations, and identified issues for further consideration. The Google document included reference to all source documents, process notes and an electronic audit trail. The trustworthiness of the process and analysis was ensured by member-checking, creating an audit trail and transparency during the process [39, 43].

The cases in this multi-case study research were four different MOOC providers with differing geopolitical locations and sizes (Table 1).

The rationale for the selection of these four MOOC providers related to: (1) locations in very different geopolitical and legislative contexts; and (2) established providers with a wide range of enrolments. In following Yin [43], the selection and analysis of these four cases is not an attempt to

Table 1. Background information for each MOOC provider

	edX	Coursera	iversity	FutureLearn
Geopolitical location	United States of America	United States of America	Germany	United Kingdom
Regulatory environment	United States of America	United States + California rules	Federal Republic of Germany	European Data Protection Act 1998
Number of courses	1750+	2500+	110+	640+
Number of students	10,000,000	25,000,000	1,000,000	6,688,892
Documents analysed	a. Terms of Service & Honor Code b. Privacy Policy	a. Terms of Use b. Privacy Policy	a. Terms of Use b. Privacy Policy	a. Accessibility and inclusion policy; Cancellation and refund policy; Cookie policy; Data protection policy; Privacy policy; Research ethics; Terms and Conditions

generalise to all MOOC providers. Case studies attempt to generalize to theoretical positions, inform theoretical development and illustrate/describe the functioning of specific theoretical concepts in chosen locations [43].

Two MOOCs were selected from the United States of America (edX and Coursera) and two from Europe (iversity and FutureLearn). The geopolitical locations allowed some consideration of whether US and European legislation may have shaped approaches to 'personal data', consent and use. The units of analyses were the providers' Terms and Conditions (or Terms of Use) and privacy statements. While all four providers have a document stipulating the terms and conditions of use and policies pertaining to privacy, edX also provides an Honor code, and FutureLearn has policies pertaining to accessibility, cookies, and research ethics.

The quantitative analysis of this research study applied text data mining by deriving categories, extracting frequently mentioned terms and making correlations. Silge and Robinson's [24] techniques and recommendations were followed for the text analytics. Information was collected and each of the privacy policies and terms of use (on 13-Sep-2017) studied, pasting content in a text file separately using UTF-8 character encoding. In total, there were eight text files, 120 pages, and 36,965 words.

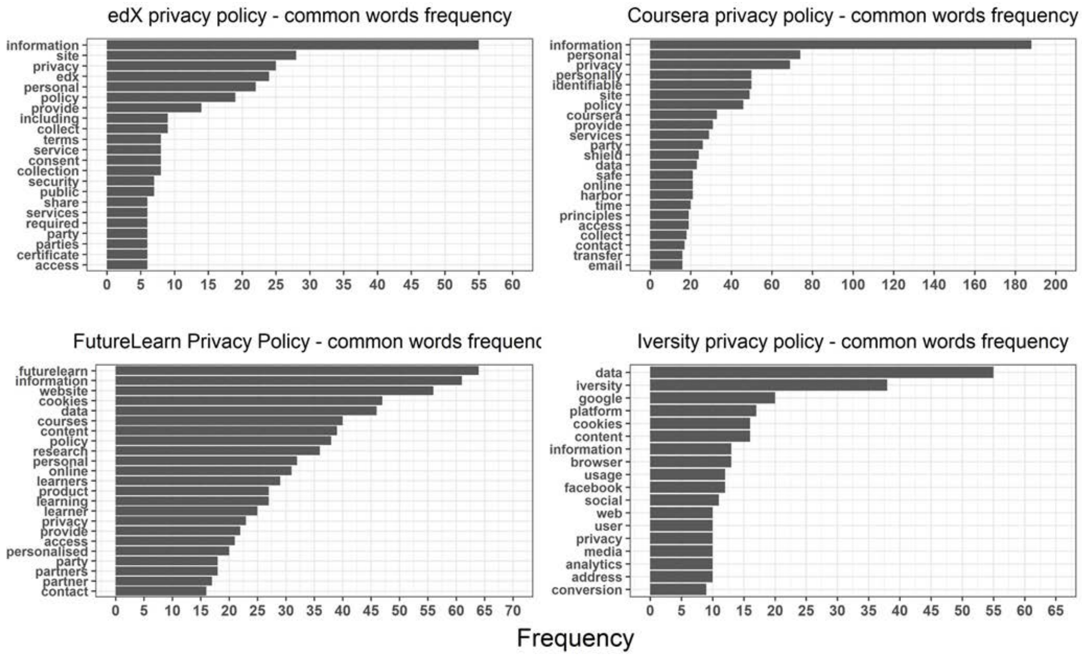


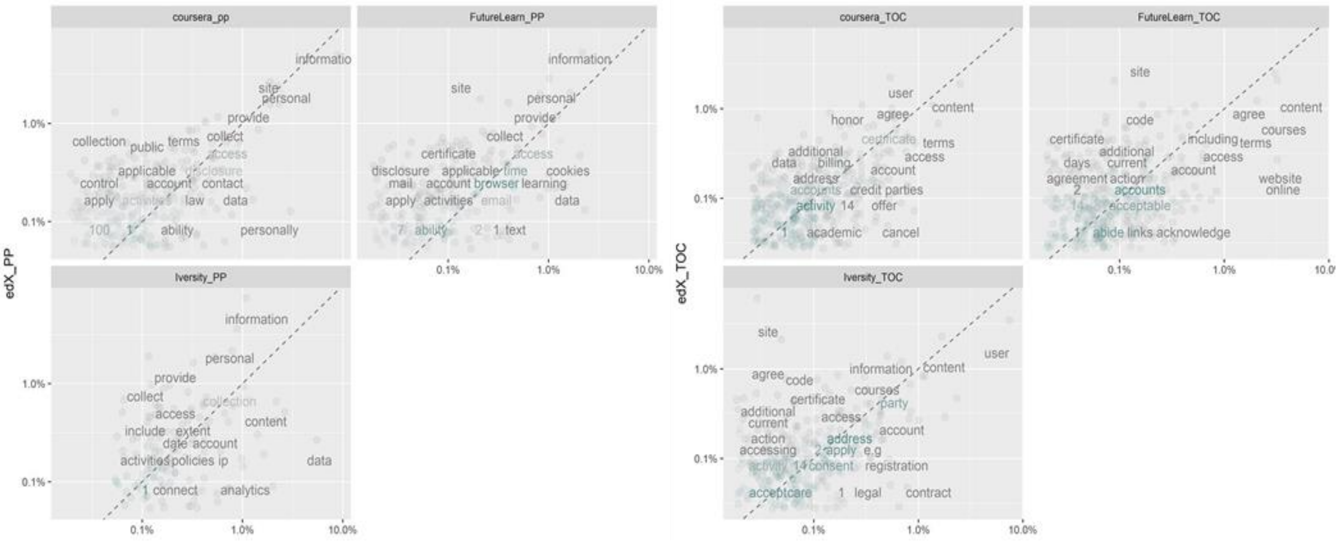
Figure 1. Common words found in the Privacy Policies of the studied 4 MOOC providers

The qualitative part of this research involved a *directed* content analysis [19] and specifically looked at definitions of, and indications of the scope of 'personal data', consent and the uses of data to inform pedagogy, assessment and learning. Within the constraints of concerns about the subjectivity of all research, including qualitative research, the researchers ensured trustworthiness of the analysis and findings by quoting from publicly available source documents, cross-checking between researchers, creating an audit trail of the selection of texts and analysis, and using critical peer checks as proposed by Rule and Vaughn [39].

Data Pre-processing

Prior to text analysis, data pre-processing was conducted. Each text file was imported as a distinct data frame to the R software (http://www.rproject.org), and each data frame converted into a tibble. A tibble is a new form of data frame by which the imported text into the computer software is retained as a string and not as any other form i.e. factor, numeric. In order to process the text, each text tibble was tokenized. The tokenization process "is the process of demarcating and possibly classifying sections of a string of input characters, and resulting tokens are passed on to some other form of processing [50]. Generic stop words such as 'will', 'be', 'about' were removed by using three popular lexicons and add-on packages such as the tidytext library (https://cran.r-project.org/web/packages/tidytext/index.html) were applied for processing and implementation. The interactive visualization of the text analytics in this research was implemented and modified to suit our particular goals using ggplot2 R library [16]. A frequency analysis of the words was carried out, followed by a comparative analysis and rank correlation. The final phase of the analysis entailed a qualitative content analysis specifically looking at the three categories of (1) personal data; (2) consent; and (3) use.

The purpose of the quantitative part of this research was to function as a tentative baseline indicator for the terms and frequency of terms these providers address issues of 'personal data', 'consent', and 'intervention'. Following Stemler [47], word frequency does not serve as sufficient



(a) Privacy Policies

(b) Terms and Conditions

Figure 2. Comparison between MOOC platforms based on frequent words proportion

evidence for the relative importance of these terms or concepts, but rather points to words as basis for further interrogation. Also see Hsieh and Shannon [19].

ANALYSIS AND FINDINGS

To recap, the research questions focus on definitions of personal data, consent options and detail around uses of student data.

Frequency and Comparative Analysis

Frequently used terms in the collected and processed texts were identified by extracting the most common words for each Privacy Policy and Terms and Conditions document. In Figure 1, the common words frequency for the 4 MOOC policies are depicted as a flipped bar plot.

The edX and Coursera platforms both have high occurrence of ‘information’, ‘personal’, ‘privacy’, and ‘policy’. Other interesting terms used by the US providers are ‘access’, ‘consent’, ‘data’, and ‘collection’. The FutureLearn Privacy Policy has ‘information’, ‘cookies’, ‘policy’, and ‘research’ as the most common terms. Other frequent phrases are ‘consent’, ‘data’, and ‘personalised’. Interestingly, policy documents for the German-based MOOC provider, iversity, had ‘Google’, ‘Facebook’, ‘data’, ‘analytics’, ‘media’ and ‘privacy’ as most frequent words.

While word frequency analysis has limitations [47], one of its advantages is to highlight high frequency words as well as words that are unique in one particular context. For example, ‘Google’ and ‘Facebook’ appear in the case of the German provider, iversity, but are absent, at least in the list of most frequently used words, from the other providers. The appearance of these words may warrant further investigation.

As might be expected, high frequency words often related to procedural issues such as usage of the websites, browsing content, payment and refunds, and registration on their platforms. However, there were a few interesting words, such as ‘photo’ in the edX platform. Further exploration of words associated with personal data, sensitive data, consent, and intervention was conducted in line with the research focus. Consent synonyms such as ‘opt in’, ‘opt out’, ‘authorize’, ‘agree’, and ‘accept’ were examined as valid proxies. Sensitive data was limited to the term itself, and intervention was related to broadly equivalent terms such as personalization, recommendation, individualization, respond, adapt, etc.

The results show that sensitive data is mentioned only by Coursera. Intervention as a term is not found in any of the privacy policies nor in the terms of use. However, edX does refer to “personalization and recommendation” as potential actions using student information. The other platforms use synonyms of intervention to reflect improvement to their websites, enhancing courses or individualizing the user experience. In respect to personal data, Coursera used ‘personal information’ and ‘personal data’ as interchangeable phrases in their Privacy Policy. Consent and equivalent terms are frequently used. In particular, ‘consent’ and ‘agree’ appear together when informing users about data collection, website’s policy, and general regulations.

Since each platform’s policies may differ from the other in length, consideration was given to a comparison of the word frequencies based on proportions (see Figure 2, which illustrates word proportion comparisons with reference to edX). In both Figures 2a and 2b, phrases close to the midline have equivalent frequencies. For instance, in both edX and Coursera Terms and Conditions, words like ‘accept’, ‘accurate’, ‘certificate’, ‘content’ or in edX and Coursera Privacy Policy documents words like ‘collect’, ‘aware’, ‘access’, ‘personal’, ‘information’. Words that are far away from the midline are prominent in one group but not the other. For example, words like ‘honor’, ‘code’, ‘arising’ are more frequent in the edX Terms and Conditions while words like ‘services’, ‘terms’, ‘cancel’ are frequent in the Coursera document. Figure 2 suggests broader similarities for word usage between the US platforms edX and Coursera Privacy policies with fewer scattered words away from the midline.

To further inspect, the correlations of word frequencies were examined across the 4 MOOC platforms. The Pearson Product Moment Correlation was used on both privacy policy and terms of use. Table 2 shows a Pearson correlation between the four studied MOOC platforms for the Terms and Conditions documents.

Table 2. The Pearson correlation coefficient (*r*) of the terms and conditions between the MOOC providers

MOOC providers	*r*
edX ↔ Coursera	0.519
edX ↔ iversity	0.449
edX ↔ FutureLearn	0.383
Coursera ↔ iversity	0.384
Coursera ↔ FutureLearn	0.470
iversity ↔ FutureLearn	0.522

p-values for all comparisons are <0.001

The strength of association between the set of words of (edX, Coursera) and (iversity, FutureLearn) can be identified as large, taking into account that the *r* value is >0.50 and the p-value <0.001. Other correlations can be identified as medium having the coefficient *r* value bigger than 0.3 and less than 0.50 with *p*-value <0.001.

Table 3. The Pearson correlation coefficient (r) of the privacy policies between the MOOC providers

MOOC providers	*r*
edX ↔ Coursera	0.902
edX ↔ iversity	0.204*
edX ↔ FutureLearn	0.529
Coursera ↔ iversity	0.263
Coursera ↔ FutureLearn	0.571
iversity ↔ FutureLearn	0.484

p-values for all comparisons are <0.001

* *p*-value < 0.02

Table 3 shows the correlation between the MOOC platforms for the Privacy Policies. It is interesting to note the correlation between the Privacy Policies of the US platforms. The *r* value is 0.902 which is high and the *p*-value <0.001 provided that generic stop words are taken out of the calculation. On the other hand, the (edX, iversity) correlation comparison is small with *r*-value <0.30 while other Privacy Policy comparisons range around the medium strength of association.

Qualitative Analysis

In following Hsieh and Shannon [19] the qualitative analysis entailed a directed content analysis informed by theoretical constructs derived from literature. The definition and use of the concepts – 'personal data', 'consent' and 'intervention' were marked, coded, and analysed. Direct quotations from the respective documents were selected to illustrate the scope and use of the concept in the context of a particular MOOC provider.

Personal data

The selected policies of the 4 MOOC providers were reviewed to examine how issues of personal and sensitive data were described. Each provided definitions of personal data and identified how it was collected (e.g., as part of registration, as part of study activities and other online activity and via direct communication between student and the MOOC provider). Definitions of personal data ranged across the platforms. All include email, name, forum posts, student generated or shared content and IP address. Student generated content can include anything shared by the student such as photos, career information, tertiary institutions attended, contact details, and personal news and comments.

EdX policies state that personal information is "any information about yourself that you may provide to us" (para 6). It may include (but is not limited to) contact information, gender, date of birth, occupation. The site also captures information around logins, webcam photos, links to social media accounts, and information about student performance and patterns of learning. Its Privacy Policy excludes information collected from students in other ways (for example, over the phone, by fax, or through conventional mail). However, although edX lists the scope of their personal data, they include a phrase that suggests that they may collect other forms of data. No further information is provided.

Coursera includes other potential items as personal data. For example, their policies include date of birth, webcam headshot, a photo identification document, sample of typing patterns, information regarding income. If logging into another site from Coursera, they also collect text and/or images of Personally Identifiable Information available from the third party site: "We may receive Personally Identifiable Information when you access or log-in to a third party site, e.g., Facebook, from our Sites. This may include the text and/or images of your Personally Identifiable Information available from the third party site" (Coursera privacy policy).

Similarly, iversity identifies potential opportunities to sweep up other information from third party sites, e.g.: "If you log onto the platform using your Facebook account (Facebook Connect), your Facebook profile and our offer will be linked and you have thereby consented to the collection, processing and use of the data contained in your public Facebook profile by iversity" (iversity privacy policy, para 1.1).

FutureLearn appears to collect less personal information than its US counterparts though does also specify gender, address and educational background. Compared to other sites, it is reasonably open. For example, the site provides a detailed breakdown of the cookies it collects *and* the purposes for which they are collected, e.g., to target users with ads on Facebook. It also makes explicit how to opt out of being tracked by Google Analytics.

Only Coursera specifically mentions sensitive data which it classifies as: race, ethnic origin, sexual orientation, political opinions, religious or philosophical beliefs, trade union membership or information that concerns health. The uses to which student data were put were also described, although not always comprehensively. On the whole, uses of student personal data were given as: improvement of course (delivery) for future cohorts; authentication (if linked to certificates of achievement); scientific/research purposes; tracking of individual/aggregate attendance, progress and completion; personalisation; marketing; third party sharing – this included administrative processing, posts within public forms, peer-peer contact and links to third party services.

Consent

Consent around uses of data remains largely unsatisfactory. Typically users consent at the time of registration, to allow the providers to collect, use, disclose and retain information. If this is not acceptable, edX users are advised "If you do not agree with these terms, then please do not access, browse, or register for the Site" (edX Privacy Policy, para 5). It is made clear that opting not to provide certain information leads to a probable inability to use the MOOC services. As Coursera is the only provider to explicitly detail sensitive data, it offers an ability to opt out of uses of personal information and to provide specific consent *for the use of sensitive data for direct marketing purposes/disclosure to a third party*, or for uses other than for the original purpose (although, again, there is no information provided on what the original uses of such sensitive information might be).

Interventions

The extent to which use of data leads to personalisation is not always made clear. It is presumed to mean presentation of different versions of course materials and software 'best suited' to the learner (for example, by assessing levels of ability). There is no expansion of tracking and whether this leads to direct intervention in cases of lack of progress, for example. What remains very much unclear are the uses to which *non-course specific* personal data might be put. None of the providers make explicit reference to this and it is perhaps of some concern that MOOC providers are gathering information relating to, e.g., Facebook likes.

DISCUSSION

The quantitative and qualitative document analysis of the 4 MOOC providers flags up several issues. For example, the quantitative analysis suggests that user data are increasingly used across platforms like Facebook and Google (e.g. in the case of iversity). This confirms research by West [45] and Pasquale [12] that point to the commercial value of data, how data are shared across platforms and how these data assemblages are increasingly algorithmically driven without human oversight, or increasingly, understanding. User data is being collected on a large scale, and seemingly unrelated to the provision of the learning service. There is apparently no intentional, specified data collection, but rather a scraping of available data [49] in a process described by Cormack [1] as looking for patterns. As Figure 1 illustrates, the use of data does not appear on the list of frequent words, while collecting appears more frequent in most of the US and European providers. It would appear that the emphasis is, indeed, on collecting a range of data without a clear sense of how that data will be used. This is in stark contrast to the proposal by the International Working Group on Data Protection in Telecommunications' [20] proposal that institutions should only collect specific information needed for a specific purpose.

The quantitative analysis shows similarities between the privacy policies for the two US providers, but suggests significant differences between the American and European-based MOOC providers. For example, Figure 1 shows that the US providers share a number of higher frequency terms, such as personal information, collection, and services. In fact, the frequent repetition of such terms strongly affects the results shown in Table 3. The high correlation of word frequencies between the edX and Coursera policies suggests significant commonalities in the ways that the two US MOOC providers privacy policies handle issues of interest here (i.e. consent, personal information, intervention,...etc.). The European MOOC providers, on the other hand, are assumed to follow the European Data Protection Act. Our quantitative analyses shows that there are variation gaps in handling such critical concepts as privacy and personal data. Initiatives such as the Privacy Shield [27] will, however, offer protection for users by extending European data protection standards far beyond the boundaries of the European Union.

Where mentioned, uses of data to personalise a learning experience are not explicit. Even if there were a credible means of enabling informed consent, it would not be easy to identify what is being consented to. In the light of the implementation of the GDPR and greater sensitivity among users regarding collection and uses of data by cross-border providers, MOOC providers will need to be much clearer with regard to the scope and purpose of data collected for it to qualify as legitimate use. The inclusion of sensitive data (on the part of Coursera) and the broad sets of personal information collected (in all 4 cases) is of concern. Although some learning analytics practitioners suggest approaches to handling sensitive data in online environments (e.g., [9, 29, 37]), the roots of learning analytics as a means to improve student outcomes appear distorted when applied to informal learning platforms such as the MOOCs reviewed here. In the next section, recent suggestions for establishing the boundaries for reasonable consent are considered.

Towards an alternative heuristic

Current approaches to handling personal data and consent are largely unsatisfactory [1, p. 1]. Traditionally consent for research purposes is sought on the basis of an already formulated hypothesis *before* data are collected and with participants informed about the implications of providing data and how it might affect them. In a big data environment there is often no prior hypothesis. The purpose for the collection and analysis is to seek patterns and then to match individuals to those patterns. It is unrealistic and contradictory in an analytics context to expect a pre-declared hypothesis and to seek consent for an unknown purpose.

Cormack [1] proposes an alternative framework for consent based on differentiating between pattern-finding and pattern-matching behaviours. In the former, the purpose is to identify patterns such as student engagement and progression. Seeking and finding patterns involves no direct risk for students as their learning is not directly affected. For example, analysis of general student engagement and progression patterns could be used to improve the next iteration of a course, changing the sequence of activities, for example. But when patterns are applied to specific individuals in what Cormack [1] calls pattern matching, consent is required. "Consent is thereby sought at the time when most information about the proposed intervention is available but least impact caused by it" (p. 5).

In the context of a specific focus on a user's perspective, the benefits for individuals include being made aware of the scope and purpose of data collection and having control over their data. Should an analysis suggest a pattern "not compatible with the purposes stated when data were collected, that pattern and insight may not be acted upon unless the individual agrees" [1, p. 7]. Individuals would have a right to provide explicit consent where the intervention will significantly alter his or her experience.

Reflecting on the implications of the forthcoming EU GDPR, Jisc [31] suggests the following three-tiered response: 1. Not asking for consent for the use of non-sensitive data for analytics (on the basis that this may be considered as of legitimate interest); 2. Asking for consent for use of sensitive data (which, under the GDPR, will be labelled "special category data"); 3. Asking for consent to take interventions directly with students on the basis of the analytics.

This implies that if the data in question are not considered 'sensitive', and do not form the basis for any intervention, *consent is not required*.

Reflecting on the Jisc guidelines, Prinsloo and Slade [37] raise concerns. In following Kitchin [41] they accept that data do not exist independently "of the ideas, instruments, practices, contexts and knowledges used to generate, process and analyse them" (p. 2). As a result, they question that consent *is not* required for non-sensitive data assumed to be of legitimate or public interest. Prinsloo and Slade [37] propose that "all data, depending on the context, might reasonably be considered to be sensitive." They also point out that what may constitute non-sensitive data in one context or at a particular time, may be considered sensitive in another context and/or time [25]. Cormack's alternative framework [1] which differentiates between initial consent for the collection of data and specific consent when data are used to intervene in the choices students have or/and in adapting their learning experience or access to resources is preferred. Although there are practical difficulties in doing so in a context increasingly shaped by a pattern seeking and algorithmic approach, an expectation that users should consent to uses of personal data unknown at the point of registration seems to be an unreasonable and unethical one.

CRITICISM AND RESPONSES

In the context of education, while issues surrounding the collection, analysis and use of student data is an emerging focus for research and practice, the issue of student consent is relatively marginal. Generally, institutions have accepted that initial consent provided by students at the point of registration as sufficient to address possible concerns regarding the subsequent collection, analysis and use of student data by a range of institutional stakeholders. While the ethical considerations in the collection, analysis and use of student data by researchers are governed by Institutional Review Boards, the operational use of student data has largely not been considered [21].

It is proposed that initial consent does not provide a blank cheque to harvest and use student personal data without considering the original context of the consent and data in order to ensure contextual integrity [17]. It is also argued that the original consent does not cover data and uses unforeseen at the point of consent. A rationale has been provided to consider that when data are required that were outside the original scope of consent, or where provided data are used for uses that were not entertained at the initial point of consent, that the alternative framework proposed by Cormack [1] opens up alternative approaches.

Current definitions in institutional Terms and Conditions of 'personal data' and consent, and statements of how the data will be used, do not fully consider the role of context in which the data are shared and in which consent is provided.

(IN)CONCLUSION

This paper provides evidence that definitions of 'personal data', and frameworks for 'consent' and 'use' differ not only between providers, but between providers in different geopolitical locations. From a user's perspective, understanding what data are included and excluded, what one's consent entails and how it will impact on one's learning experience and, increasingly downstream [22, 26], are of growing importance. In the context of the MOOC providers studied, there is no evidence that student data is used to increase success and retention, or to offer individualized support. Despite the frequency of words like 'privacy' in the analysed documents, it seems that consent is of little or no consequence and, indeed, is unbearably light.

Limitations of the Study

This research focused on four MOOC providers from two specific global contexts. We acknowledge that the multi-case study research design does not allow us to generalize to populations or universes, but to theoretical propositions. We furthermore acknowledge that the emphasis on a 'user's perspective' limits the research. However, the quantitative analysis of this study has built on a tokenization of individual terms (i.e., uni-gram). There was no analysis of the content of the privacy policies nor the terms of use on consecutive sequences of phrases (i.e., n-grams). For the purposes of this study, a lay-person's legal perspective was taken.

ACKNOWLEDGEMENTS

We thank the reviewers for their constructive feedback, comments, and suggestions for improvement.

REFERENCES

[1] Andrew Nicholas Cormack. 2016. Downstream Consent: A Better Legal Frame- work for Big Data. *Journal of Information Rights, Policy and Practice* 1, 1.

[2] Bart Rienties, Avinash Boroowa, Simon Cross, Chris Kubiak, Kevin Mayles, and Sam Murphy. 2016. Analytics4Action Evaluation Framework: A Review of Evidence-Based Learning Analytics Interventions at the Open University UK. *Journal of Interactive Media in Education* 2016, 1.

[3] Charlotte Strange. 2011. *Privacy concern and student engagement in the virtual classroom*. Technical Report. University of Victoria. 73 pages.

[4] Chris Gilliard. 2016. Digital Redlining, Access and Privacy.

[5] Daniel J. Solove. 2008. *Understanding Privacy* (2/28/10 edition ed.). Harvard University Press, Cambridge, Massachusetts.

[6] Daniel J. Solove. 2013. Introduction: Privacy Self-Management and the Consent Dilemma. *Harvard Law Review* 126, 1880-1903.

[7] David Beer. 2017. The social power of algorithms. *Information, Communication & Society* 20, 1.

[8] Dragan Gasevic, Shane Dawson, and George Siemens. 2015. Let’s not forget: Learning analytics are about learning. *TechTrends* 59, 1 (Jan. 2015), 64-71.

[9] Elaine Sedenberg and Anna Lauren Hoffmann. 2016. *Recovering the History of Informed Consent for Data Science and Internet Industry Research Ethics*. SSRN Scholarly Paper ID 2837585. Social Science Research Network, Rochester, NY.

[10] Ewa Luger, Stuart Moran, and Tom Rodden. 2013. Consent for All: Revealing the Hidden Complexity of Terms and Conditions. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’13), 2687-2696.

[11] Ewa Luger, Tom Rodden, Marina Jirotka, and Lilian Edwards. 2014. How do you solve a problem like consent? - The workshop.

[12] Frank Pasquale. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, Cambridge, MA, USA.

[13] Gary T. Marx. 2016. *Windows into the Soul: Surveillance and Society in an Age of High Technology*. Univ of Chicago Press.

[14] Gary Thomas. 2011. *How to do your case study: A guide for students and researchers*. Thousand Oaks, CA: Sage Publications, California.

[15] Gert Biesta. 2007. Why “What Works” Won’t Work: Evidence-Based Practice and the Democratic Deficit in Educational Research. *Educational Theory* 57, 1, 1-22.

[16] Hadley Wickham. 2009. *ggplot2 - Elegant Graphics for Data Analysis*. Springer.

[17] Helen Nissenbaum. 2009. Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press.

[18] Helen Nissenbaum. 2011. A contextual approach to privacy online. *Daedalus*, 140(4), 32-48.

[19] Hsieh, H-F., & Shannon, S.E. 2005. Three approaches to qualitative content analysis, *Qual Health Res*, 15(9), 1277-1288.

[20] International Working Group on Data Protection in Telecommunications. 2017. Working Paper on E-Learning Platforms. Washington D.C., USA.

[21] James E. Willis, Sharon Slade, and Paul Prinsloo. 2016. Ethical oversight of student data in learning analytics: a typology derived from a cross-continental, cross-institutional perspective. *Educational Technology Research and Development* 64, 5.

[22] Jeffrey Johnson. 2017. Structural Justice in Student Analytics, or, the Silence of the Bunnies – The Other Jeff. Philadelphia.

[23] Joshua A. T. Fairfield. 2017. *Owned: Property, Privacy, and the New Digital Serfdom*. Cambridge University Press, UK.

[24] Julia Silge and David Robinson. 2017. *Text Mining with R: A Tidy Approach*. O’Reilly Media, Inc.

[25] Linnet Taylor. 2017. What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, (Jul–Dec 2017),1-14

[26] Manuela Ekowo and Iris Palmer. 2016. *The Promise and Peril of Predictive Analytics in Higher Education: a Landscape Analysis*. Technical Report. New America.

[27] Mark Scott. 2017. Regulators without borders. http://www.politico.eu/article/data-privacy-shield-regulators-without-borders/

[28] Michael Zimmer. 2008. Privacy on planet Google: Using the theory of contextual integrity to clarify the privacy threats of Google's quest for the perfect search engine. *J. Bus. & Tech. L.*, 3, 109.

[29] Mohammad Khalil and Martin Ebner. 2016. De-identification in Learning Analytics. *Journal of Learning Analytics* 3, 1, 129-138.

[30] Mohammad Khalil, Behnam Taraghi, and Martin Ebner. 2016. Engaging Learning Analytics in MOOCS: the good, the bad, and the ugly. In *Intl Conf on Education and New Developments*, 3-7, Slovenia.

[31] Niall Sclater. 2017. Consent and the GDPR: what approaches are universities taking? Effective Learning Analytics. (2017).

[32] Niall Sclater. 2017. Consent for learning analytics: some practical guidance for institutions.

[33] Paul Dourish. 2016. Algorithms and their others: Algorithmic culture in context. *Big Data & Society* 3, 2

[34] Paul Henman. 2004. Targeted! Population segmentation, electronic surveillance and governing the unemployed in Australia. *International Sociology* 19, 2.

[35] Paul Prinsloo and Sharon Slade. 2015. Student Privacy Self-management: Implications for Learning Analytics. In *Proc of 5th Intl Conference on Learning Analytics And Knowledge*. 83-92.

[36] Paul Prinsloo and Sharon Slade. 2017. An elephant in the learning analytics room: the obligation to act. In *Proc of 7th Intl Learning Analytics & Knowledge Conference*. ACM.

[37] Paul Prinsloo and Sharon Slade. 2018. Student consent in learning analytics: the devil in the details? In J. Lester, C. Klein, H. Rangwala, and A. Johri (Eds), *Learning analytics in higher education: Current innovations, future potential, and practical applications*. Accepted. In press.

[38] Petar Jandrić, Jeremy Knox, Hamish Macleod, and Christine Sinclair. 2017. Learning in the age of algorithmic cultures. *E-Learning and Digital Media*.

[39] Peter Rule and John Vaughn. 2011. *Your guide to case study research*. Van Schaik Pretoria.

[40] Rebecca Ferguson and Doug Clow. 2017. Where is the Evidence?: A Call to Action for Learning Analytics. In *Proc of 7th Intl Learning Analytics & Knowledge Conference*. 56-65.

[41] Rob Kitchin. 2014. *The data revolution*. SAGE Publications Inc, London, UK.

[42] Robert Payne. 2014. Frictionless Sharing and Digital Promiscuity. *Communication and Critical/Cultural Studies* 11, 2 (April 2014), 85-102.

[43] Robert Yin. 2013. *Case study research: Design and methods* (4th ed.). SAGE.

[44] Robin Fretwell Wilson. 2017. *The Promise of Informed Consent*. SSRN Scholarly Paper ID 2913863. Social Science Research Network, Rochester, NY.

[45] Sarah Myers West. 2017. Data Capitalism: Redefining the Logics of Surveillance and Privacy. *Business & Society*, https://doi.org/10.1177/0007650317718185.

[46] Satu Elo and Helvi Kyngäs. 2008. The qualitative content analysis process. *Journal of Advanced Nursing* 62, 1, 107–115.

[47] Steve Stemler. 2001. An overview of content analysis. Practical assessment, *research & evaluation*, 7(17), 137-146.

[48] Tore Hoel and Weiqin Chen. 2016. Implications of the European data protection regulations for learning analytics design. In *The International Workshop on Learning Analytics and Educational Data Mining*. Kanazawa, Japan.

[49] Walter Frick. 2017. Do Tech Companies Really Need All That User Data. Harvard Business Review.

[50] Wikipedia. 2017. https://en.wikipedia.org/wiki/Lexical_ analysis.

[51] Wilfried Bos and Christian Tarnai. 1999. Content analysis in empirical social research. *Intl J. of Educ Res* 31, 8, 659-671.