Ethical Oversight of Predictive Analytics in Higher Education:

A Case Study

Stephanie L. Kane / William B. Davis

Washington State University

June 2026

Keywords

Predictive analytics, machine learning, student success, ethical AI, higher education

Abstract

The use of machine learning (ML) or artificial intelligence (AI) algorithms to make decisions about persons has brought new opportunities and new challenges. While these tools can increase efficiency or lead to new insights, algorithms deployed in human systems do not necessarily reflect the full complexity, nuances, and ethical frameworks of those systems. In many cases, the underlying training data may be limited, or including sensitive information, eliciting privacy concerns. Furthermore, deployment of these models and conveying the results to the stakeholders involved presents challenges and requires appropriate transparency. In recent years, the use of AI/ML-based predictive analytics in higher education has boomed. Institutions using these systems should establish appropriate oversight for their use and deployment. Effective oversight includes not only dissemination of model performance and accuracy, but also how to implement the tool and communicate to stakeholders about results. Oversight of this type of work should not be left to developers or a few technical people, but should be a concerted effort across a diverse committee with a range of expertise and perspectives. In this paper we present a case study as a model for effective ethical oversight that addresses many of these concerns.

Corresponding Author: Stephanie Kane, Washington State University, slkane@wsu.edu To quote this article: Kane, Stephanie L. and William B. Davis, 2026. "Ethical Oversight of Predictive Analytics in Higher Education: A Case Study" *Journal of Ethics in Higher Education* 8.2(2026): 1–28. DOI: https://doi.org/10.26034/fr.jehe.2026.9497

1. Introduction

Over the past few decades, advancements in computing speed and data storage capabilities have resulted in the ability to create, store, and analyse vast quantities of data. Analytical tools and algorithms, such as those used in machine-learning (ML) and artificial intelligence (AI) systems are now applied across many areas of modern society, consuming both structured and unstructured data for the purpose of forecasting, prediction, and decision support, including to make decisions for and about people.

Sociotechnical systems, where social, economic, and/or cultural considerations are important, require additional assessment throughout the development and deployment process, moving beyond model accuracy and predictive power to also consider questions of privacy, fairness, and respect for persons. Decisions made by these algorithms can have profound economic and social consequences at both the individual and population levels. (O'Neil 2017; Olteanu et al. 2019; Selbst et al. 2019; Rubel/Casto/Pham 2020; Birhane 2021; Schwartz et al. 2022; Broussard 2023; Abassi 2025). The greater the potential risk or adverse effects to persons or society, the higher the standard of care is necessary to ensure safety (Martens 2022). It is not enough to be assured of accuracy: when decisions are made about people, there must be both a justification for the tool use, and a level of transparency about the project that is in proportion to the implications of the decision (Olhede/Wolfe 2018; Rubel 2020).

One sector that has experienced a recent and rapid increase in the number of AI/ML algorithms deployed is higher education. Enterprise data systems, such as student information systems (SIS), learning management systems (LMS), and customer relationship management (CRM) tools, which are a necessary component for managing university business operations, generate vast quantities of data and metadata resulting in what has been termed the "datafication" of higher education (D Florea/S Florea 2020; Komljenovic/Sellar/Birch 2025). Institutions may analyse data at hand to determine how best to improve student outcomes, evaluate program efficacy, or leverage institutional resources, with examples in the realms of recruitment and admissions, financial aid awarding, student success/evaluation, and

alumni giving (Ekowo/Palmer 2016; Lawson et al. 2016; May/Iksal/Usener 2016; Bird et al. 2022; Burd 2024; Weinberg 2024). Using institutional data from an SIS or CRM, these tools can be used to make decisions or predictions at the level of an individual student. In other words, institutions are increasingly making use of large swaths of data about and from students to not only assist with evaluation and policy setting, but also to make student-level predictions on a variety of settings, including the likelihood of enrolling or being retained at the institution. The use of this type of data for analytics and assessment has opened new questions about ethics, particularly with respect to whether these types of tools balance beneficence (i.e. working towards students' benefit) with any potential harms or the loss of student agency, autonomy, and privacy (May/Iksal/Usener 2016; Corrin 2021; O'Donoghue 2023).

One complexity in sociotechnical systems (not unique to higher education) is that the majority of the AI/ML tools in use are privately developed. These products are sold to institutions by companies and consulting firms promising efficiency and accuracy of their ML/AI tools, yet these same companies typically provide limited information on the algorithm used, with models and decision making considering a proprietary process. This lack of transparency often renders these tools a "black box" for an adopting institution (Mittelstadt et al. 2016; Ekowo/Palmer 2017; Burd 2024; Weinberg 2024). Compounding the issue, developers at private companies may lack sufficient domain or institution-specific knowledge to understand the nature of the underlying data; indeed, many university administrators implementing these tools may be sufficiently removed from the data generation process to understand data limitations or constraints (Mathies 2018). The tools are often sold as a "one size fits all" solution that does not take into account the local data structure or social framework ("the portability trap") (Olteanu et al. 2019; Selbst et al. 2019;). Human oversight of these types of algorithms is often insufficient and can inadvertently increase the risk of bias and misuse (D Florea/S Florea 2020; Green 2022).

Even when models and algorithms are developed in-house, few institutions have policies or procedures governing oversight of these tools, or even policies that clearly state how and when student data may be used in

assessment (Komljenovic/Sellar/Birch 2025). A recent review of data privacy policies at 151 private and public universities in the United States found most privacy policies are unequipped for the types of data generated at a modern post-secondary institution (Brown/Klein 2020). Among their key findings: while privacy policies may outwardly grant students authority over their data, student agency is quite limited given the many sources and types of data. Additionally, privacy policies treat data as static and restricted to traditional educational records (e.g. grades, transcripts), rather than broadening policies to include data created by IP addresses, card swipes, and/or metadata within an LMS. When data are consumed or analysed by third party vendors, the data use agreement is a service level agreement between the vendor and the institution, with inconsistent protections given to students (Reidenberg et al. 2013; Reidenberg/Schaub 2018). Lastly, privacy policies do not often articulate the "why" for a particular project or define a particular project's rationale, risks, and benefits (Brown/Klein 2020).

When implemented, institutions do not often give sufficient forethought to effective deployment of the tool, which can make all the difference in the success of a project. Those responsible for administering the tool may lack sufficient training, or the project may lack clear rubrics for how to interpret the results and communicate to students or other stakeholders. Models may be deployed in an inappropriate setting, generate errors, raise privacy concerns, and/or poorly understood. This oversight can result in an outcome that is the opposite of what was intended. During the deployment of a student early warning indicator in Australia, while the bulk of communications sent because of the model output were found to increase student retention, some academic staff sent email communications to students threatening them with failure; these messages were found to be very demotivating for students (Lawson et al. 2016). This example demonstrates that an AI/ML product implemented without clear institutional oversight and user training can lead to tools that increase harm, rather than mitigate it.

In this paper, we outline our institution's deployment and oversight of a ML algorithm developed in-house to assist in student success and retention. This type of model has been used extensively in higher education (Lawson et al. 2016; Bird et al. 2022; Cardona et al. 2023). We used the term "student

support model" for our algorithm. While "student risk model" is often seen in the literature, the word "risk" implies a deficit-framing perspective, and the goal of the algorithm is to increase student success rates (i.e. earning degrees). The purpose of the model used at our institution was to serve as an early warning indicator, notifying advisors or other stakeholders as soon as possible in the first semester of enrolment that a student is at risk of not being retained in the first year.

The paper is organized as follows. First, we describe the general motivation and framework for ethical oversight that shaped our institution's decision making around model deployment and evaluation. This section includes a brief review of the literature on ethical deployment of sociotechnical systems, as well as the current state of the regulatory landscape. In section three, we give an overview of the model, including design decisions informed by concerns regarding algorithmic fairness. We also describe the process to establish ethical oversight and evaluation of the tool. In section four, we describe the results of our initial assessment. The paper concludes with a discussion of how we are using the process in an iterative manner to continue to improve both the oversight and usability of the tool, as well as some project limitations. It is hoped that the case study presented in this paper could serve as a model to other institutes of higher education who seek to deploy ML/AI algorithms while maintaining ethical oversight and respect for persons.

2. Framework for Ethical Oversight

In classical Aristotelian ethics, the goal of the moral person is to optimize between two extremes: deficiency and excess. When this objective is applied to algorithms, deficiency results from ignoring all available data and applying no model at all, whereas excess would be the use of all available data, without regard to privacy, fairness, or potential misuse or misapplication (Martens 2020). In the context of student success, looking at univariate or bivariate relationships is often insufficient for examining risk factors leading to poor grades or non-retention at the institution. ML/AI has the potential to find patterns and make predictions at scale and with high accuracy. Thus, the moral position would be to seek to optimize between these extremes and

deploy an algorithm that carefully considers the available data and any data limitations, balances accuracy and fairness in its predictions, is subject to sufficient human oversight, and is employed in a manner that maintains respect to persons (Olteanu et al 2019, Rubel/Castro/Pham 2020).

Unfortunately, little guidance exists for institutions to reference; legislation and policy in the United States lags some of its peer nations in this arena. Existing regulations, such as the Family Educational Rights and Privacy Act (FERPA), are inadequate for fully addressing issues raised by the analytical tools currently available (Brown/Klein 2020; O'Donoghue 2023). Due to both historical constructs and case law precedence, the United States does not grant the same level of rights to persons in the arena of data privacy as the European Union (EU) (Jones 2017). While some states, such as California, are beginning to create laws around data privacy or AI decision making, organizations and agencies have little legal guidance to work with when assessing algorithms. A recent United States Government Accountability Report raised this issue to the House Subcommittee on Research and Technology, Committee on Science, Space, and Technology, specifically calling out higher education as one of the areas in which these algorithms could use additional oversight (GAO 2022). Currently, a gap exists between the capabilities of ML/AI and the ability of policy makers and education professionals to use, assess and implement these tools.

Despite lacking clear legal frameworks (or even informal guidelines) within the U.S., other countries have developed more robust policies or recommendations to guide the development and deployment of algorithms used in sociotechnical contexts. One such formal structure, developed in the United Kingdom (U.K.) is the Data Ethics Framework (Drew 2018; CDDO 2020). While not legally binding in the same way as the EU's General Data Protection Regulation (GDPR) (GDPR 2016), this rubric offers a clear path to assessing the ethical rigor of a sociotechnical tool. Other researchers have created rubrics to define each of the steps along the pathway to develop and deploy ML/AI tools in sociotechnical systems. Martens (2022) recommends a process dubbed the "FAT" method: Fair, Accountable, and Transparent. In this process, the five key steps are 1) data gathering, 2) data pre-processing, 3) modelling, 4) evaluation, and 5) deployment. Fairness is primarily

assessed during the first three stages, while transparency is assessed during the last three stages. Accountability should be in place throughout the process. An alternate design was described by Drew (2018) in her review of the U.K. Data Ethics Framework. This approach uses a service design approach with five steps and differs from the FAT method in that the first step is to bring awareness to the problem to be solved and communicate how ML/AI may be useful in the specific context. The initial step also includes a clear articulation of the benefit to public good and could be described as "defining the problem statement." Ideally, the work would not be performed by one or two individuals, but a team featuring individuals with different perspectives.

Frameworks such as these are useful for outlining the steps in project design. In practice, stages often blur into one another, and a good system will have feedback and evaluation at all stages. Indeed, the rapidly evolving nature of development in AI/ML and other technological advances demands a dynamic and agile response (Abassi 2025). In this paper, we reconcile these minor differences among the frameworks as follows: 1) define the problem statement and the rationale for using AI/ML within the specific context of higher education, 2) gather and assess the available data, including examining potential limitations of the data or privacy concerns, 3) develop the objective function and model, 4) evaluate the model for accuracy and fairness, and 5) deploy the model with appropriate communication and monitoring systems in place (Figure 1).

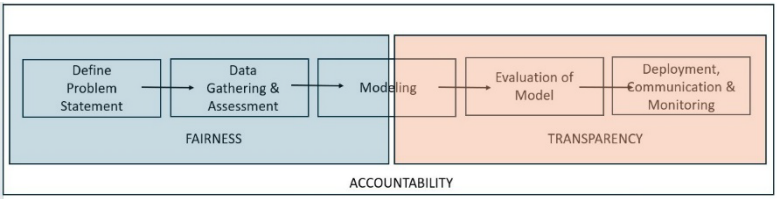

Figure 1: The Ethical Data Science Process. Adapted from (Drew, 2018; Martens, 2022)

Using this approach, the first step is to define the problem and the need for an AL/ML algorithm. In their papers, Drew (2018) and Green (2022) recommend involving the public to make an ethical determination of the need

for the project. In practice "the public" may be a much narrower group of stakeholders; stakeholders within a student success context will include senior administrators, faculty, advisors, financial aid administrators, student affairs staff, technical staff, and students (Abassi 2025). Clearly defining the problem and the use case for AI/ML will help to achieve trust and buy-in at all stages. Early involvement of end users will also ensure that the results can be easily obtained and applied.

The next step in the process is data gathering and assessment. Models used in sociotechnical systems must pay careful consideration to the training data used and determine any data sensitivities or privacy concerns. In their book Martens (2022) describes three types of data that could be used in algorithmic decision making that warrant special attention: personal data, sensitive data, and behavioural data. Personal data is information that can be directly or indirectly linked to a specific individual, such as an identification number, name, or birthdate (Mittelstadt et al. 2016; Martens 2022). Sensitive data includes information which may not be unique to an individual but could have significant privacy implications if disclosed in a way that is tied to an individual, such as financial information, health or biometric data, geographic information, or demographics. This type of data may also include item types protected by laws, such as FERPA or HIPPA. Lastly, behavioural data is information tied to behaviours of persons (Martens 2020) such as visits to a website or physical visits monitored through GPS or card swipes. Within the context of educational records, all three of these domains may be present, making models used in these settings particularly sensitive. Determining what data could and should be used in a particular project and how privacy of individuals is maintained is a key feature of an ethical design process.

Another issue that must be considered at this stage of the project is what, if any, data limitations may exist. Data limitations or concerns can be grouped into three main categories (1) statistical, which includes directional errors (bias), non-representative data, label issues, and model fairness, (2) systemic constraints (sometimes referred to as structural or historical constraints), and (3) human error, including decisions in both model development and deployment (Mittelstadt et al. 2016; Mehrabi et al. 2021; Fountain 2022; Schwartz et al. 2022; van Giffen/Herhausen/Fahse 2022).

Statistical bias can be defined as the difference between the expected and true values of the function and is often directional (Bain/Engelhardt 1992). Importantly, statistical bias can occur for the entire population, or it may only occur on a subset of the observations. In other words, the estimator may function well for the population as a whole, but function poorly (and be directionally biased) for a subset of the observations.

Systemic or structural constraints may be a factor in data limitations. Observations may be clustered, whether geographically or environmentally. Our institution includes six campuses, and degree programs and resources vary across campuses. Students attending the same campus would have experienced a similar learning environment that may be different from that of other campuses. When this type of structure exists within the data, different statistical techniques may be required to handle the non-independence of observations.

Human error is simply a way of stating that people make decisions at every stage of model development, and those decisions with diverse viewpoints have impacts. In their paper, Green (2022) demonstrates that individuals are often poor judges of algorithm quality and/or when human decisions should supersede those made by the algorithm. Ethical project oversight ideally involves multiple people thinking through key decision points, such as the objective function to be modelled, the algorithm used, and the accuracy and fairness assessments employed.

The next stages of the process are modelling and evaluation of the model. While described as two different stages in the framework, the model development and assessment should be iterative, and should not be thought of as a "one and done process," but rather should be assessed repeatedly over time as new training data and new observations occur. Key decisions at these stages of the project include determining what objective function should be modelled and determining the population(s) upon which the model will work optimally. In an ethically designed sociotechnical system, it is important for humans not to cede agency to the tool and instead establish processes for feedback and refinement if errors are found (Mittelstadt et al. 2016; Schonberger 2019; Rubel/Castro/Pham 2020; Birhane 2021).

Finally, even if a model has been extensively evaluated and tested for accuracy and fairness, there must be communication and monitoring during deployment. Complex tools require many decision points which may neither be inherently 'good' or 'bad' but could have unforeseen downstream impacts. Often overlooked, but of particular importance in implementing ML/AI in sociotechnical systems is deployment bias (Selbst et al. 2019; Green 2022). Deployment bias, as defined by van Giffen/Herhausen/Fahse (2022), can occur when the model is applied in a different context than that for which it was developed, or (importantly) the results are interpreted incorrectly or inappropriately by humans in a manner not supported or justified by the data and algorithm used. Mitigating this source of bias requires both continual monitoring and proper training for the humans involved in the deployment and use of the model. Best practice requires all stakeholders, from senior administrators to students, are able to view the recommendation made by the tool, have a reasonable understanding of how those recommendations were determined, and be able to identify and correct any errors (Reidenberg/Schaub 2018; O'Donoghue 2023). Taken in sum, if not identified and mitigated, decisions made at many stages of model development and deployment can create impact model utility, fairness, and overall accuracy, even as the tool is perceived as more objective than decisions made by humans alone (Ekowa/Palmer 2016; Benjamin 2019).

3. Methodological Approach

Model Development

This work took place at a multi-campus, public R1 doctoral granting institution (Carnegie Classification: doctoral institution, very high research activity). In the 2021-2022 academic year, the Office of Institutional Research, under the direction of the Provost's Office and the Academic Success and Career Center, developed and implemented a student support model to support retention goals for undergraduates across six campuses. Throughout the process, our team used the framework described in Figure 1 and implemented a service design approach (Kimbell 2011; Drew 2018). In

this framework, the design and deployment of a tool is relational and collaborative. At our institution, that meant receiving feedback from stakeholders (primarily but not limited to the advising community) in order to ensure that the output and results were clear, actionable, and easy to access.

Prior work in the area of predictive modelling and student success at the institution had revealed several issues related to data limitations and student subpopulations. First, the six campuses differ significantly with respect to their student populations, degree programs, student support services, and the amount and type of data available. The majority of students enrolling on the main campus begin as first-time first year students (FTFY) directly following high school graduation. The main campus is residential, and FTFY students have an on-campus living requirement in their first year. This campus offers a full suite of student organizations and activities, including athletics, a robust Greek system, and recreational facilities. Two of the other campuses are smaller, do not have on campus residential housing, and have an equal mix of traditional FTFY students and transfer students (students with prior post-secondary coursework following high school graduation). One campus is entirely online, with a high percentage of non-traditional students, and two of the campuses enroll only transfer students and have a more limited suite of degree program offerings. While Greek affiliation and other types of student involvement have been shown to promote student engagement and retention (DeBard/Sacks 2011; Biddix/Singer/Aslinger 2018), as noted above the availability of programming (and thus data demonstrating student involvement and connectivity) differs substantially among campuses. Furthermore, even within a campus, different student populations (e.g. traditional FTFY students vs transfer students, residential students vs commuter students) have very different academic trajectories at the institution and different factors impacting their success.

Due to these structural constraints, one key early design decision was to create separate models for each of the campuses. The main residential campus hosts the majority of the undergraduate student population and prior work had shown that any single model predicting student success across the entire population of students was likely to be skewed heavily towards factors

impacting student success on the main campus. In other words, high model accuracy would mainly be achieved by high accuracy at only one campus.

Another issue discovered in an earlier tool used at the institution was that if all students were including in a single model without taking into account their year of study or admission type (e.g. FTFY vs transfer), the model would predict that students nearing the end of their academic career were more likely to graduate than students in the first year. This result proved unworkable for advisors, as nearly all their new entering FTFY students were flagged to be at-risk of dropping out. This finding illustrates a survivorship issue within the data: students who have been enrolled and successfully certified in their major represent students who have already demonstrated the ability to succeed academically at the institution. Again, in considering an ethical framework for model development and use, a decision was made to place students into cohorts and fit models by cohort and admissions classifications. Thus, student populations were once again split. At each campus, two models would be deployed annually, one for FTFY and a second model for transfer students. While it may seem excessive to run models separately by campus and admission status, our initial work found that student status (e.g. FTFY, transfer, campus of enrolment) impacted retention, and (importantly) data availability, so much so that it was important to model each subpopulation independently.

We carefully considered variables inclusion in the model. Preliminary statistical models examining the outcomes of first-year retention and graduation helped to establish a framework for inclusion. Those initials models were logistic regression and included a suite of variables related to student demographics, geographic location, academic performance (high school GPA and standardized test scores), first-generation status, and financial information (unmet financial need). From those early studies, we learned that high school GPA, first-generation status, and unmet financial need were significant predictors of first-year retention. Low high school GPA, being a first-generation college student, and unmet financial need of more than $7,000 were found to be significant predictors of failure to be retained after the first year.

Building on that earlier work, the initial ML model used a rolling time window of five years of training data of entering students and gradient boosted classifier using decision trees. The Day Zero model, which was the model used at the start of the semester, included basic information obtained from the students' application, such as demographic variables, high school GPA, and geographic information, as well as financial aid information (unmet financial need). In addition, socio-demographic data from U.S. Census Bureau, such as poverty, income, demographic information, and education level, were included by mapping to the ZIP code of the student's last school attended (either high school or prior post-secondary institution). As the semester progressed, information from the SIS (PeopleSoft Campus Solutions) was brought in as it became available, including data on course registrations, course add and drops, holds or other service indicators, program of study, midterm and final grades, and class schedule information. As some variables were added, others were dropped as they became less relevant as predictors. With the exception of the Census Bureau data used in the Day Zero model, all information (including information from the application or financial aid) was obtained from the SIS (hereafter referred to as "administrative data").

The model consumed data nightly from the SIS, updated predictions, and exported the scores and other basic academic information on a dashboard that was visible to advisors with the advising area of the SIS. This dashboard was filterable, so that advisors might be easily able to find their own students. When viewing a particular student's data, a radio dial showed the current support score (where a high score indicated a concern the student may be in need of additional supports/was at risk of non-retention). The dashboard also contained relevant information about the student's current class registrations and grades, as well as a list of the factors that were determined to be contributing to the particular student's score. The development team met with the advising community to present the dashboard and discuss best practices for reading and interpreting the results.

Project Oversight

In order to establish oversight for this model and to develop a pilot program more broadly that could be used for other ML/AI tools at the institution, the Office of Institutional Research partnered with the Provost's Office to create a committee charged with assessment and ethical oversight of the model. The committee was first established at the start of the 2022-2023 academic year. The committee was co-chaired by a Vice Provost and the Assistant Director of Institutional Research (the authors of this paper). Initial invitees included representatives from the academic advising community, undergraduate admissions, ITS, the Provost's Office, Student Affairs, Institutional Research, and campus and college representation (originally 13 members, including the two co-chairs). All the members were selected for their expertise in the fields of academic advising, student support services, equity and diversity, instruction, and admissions. Some of the committee members had experience with earlier student success models implemented by the institution. The charge for the committee came from the Provost's Office and included authority to "review the model, help guide its maturation, and oversee the introduction and use of the student success model throughout the system." The meeting cadence was every three to four weeks, and the co-chairs met prior to the committee meeting to establish the agenda. A Microsoft Teams site was established for the committee to share documents and communicate, and administrative support was provided by the Provost's Office to assist with managing the Teams site, posting the agenda, meeting minutes, and scheduling.

The first meeting took place in October 2022 and included introductions and a broad overview of the history of predictive analytics in the student success framework at the institution. The committee also discussed some of the opportunities and challenges inherent in using these types of models within higher education. Prior to the meeting, articles had been sent to the committee to help frame the conversation. These papers included Green (2022) and Bauman (2022). As the meetings progressed, the discussion turned towards the nature and types of data included in the model. The committee generally felt that the administrative data currently included had 1) high specificity and relevance to the task, and 2) was low risk for unintended consequences. At a

subsequent meeting, the committee was assigned Bird et al. (2022) as reading material. That paper led the committee to express significant interest in using data from the LMS at our institution (Canvas). LMS data has some drawbacks, as detailed information on timely assignment submission and grades requires faculty to set up their LMS section in consistent manners, which is challenging at such a large and diverse institution. However, the Bird et al. (2022) paper recommended a more basic approach, using log-on and time spent within the system that is summed across courses, which is more feasible. We also discussed the potential of using certain student affairs data, such as card swipe information for the Student Recreation Center to gauge campus engagement, or card swipe data at academic support centres, such as the Math Tutoring Center. Some types of available data, such as card swipes into university housing, were deemed to have significant privacy implications and were removed from use consideration.

Overall, committee support was strong for finding additional sources of data for inclusion in the model. Part of this motivation was to increase the accuracy of the model in the first few weeks of the fall term. One of the limitations of the initial model was that its predictive capability in the first few weeks was low and then increased and remained high immediately following the release of midterm grades. However, by the time midterm grades are provided mid-semester, many students may already be in significant academic distress. Thus, having earlier information, such as engagement within the LMS system, or proxy variables, such as campus or academic support engagement data, could help serve as early warning indicators prior to when the first grades are posted.

Formal Project Assessment

The committee was critical for identifying potential sources of data in their respective areas, such as student organizations or academic support, as well as how data across the system could be better aligned or improved for future use in the model. The committee was also instrumental in setting privacy guardrails and determining what data might be a bit too invasive or have too many privacy concerns for our institution's level of comfort (e.g. card swipes at residence halls). However, the focus on data was narrower than the original

charge for the committee intended. Tto more broadly examine not only specific data elements, but also the implementation and use of the model as a whole, the committee co-chairs decided to have the members of the committee formally evaluate the model and implementation using a voting strategy. The rubric for assessment was largely built on the rubric suggested by the Government of the United Kingdom (CDDO 2020). The UK rubric was adapted for a higher education framework (where the "public" is internal to the institution, namely students and employees), and our particular institution's culture.

Rubric	
Theme	Description
Transparency	Information about the project, its methods, and outcomes is publicly available
Accountability	Mechanisms for scrutiny, governance, or peer review for the project have been established
Fairness	The project promotes just and equitable outcomes with negligible detrimental effects

Table 1: Accountability Rubrics for Assessment

Questions were administered during a committee meeting via a Zoom poll. The first set of questions assessed the transparency, accountability, and fairness of the project, (see Table 1). For each of these questions, the model and implementation were evaluated within the context of the institution, thus "public" refers to students and employees of the institution and not the public at large. The second set of questions focused on specific actions involved in the oversight of the model and its used, (Table 2). Each question had a five-point scale: strongly agree (SA), agree (A), neutral (N), disagree (D), and strongly disagree (SD). Ten committee members attended this meeting, though the co-chairs and the model developer abstained from the voting (seven members voting). One of the co-chairs (Kane) took notes on the discussion. Meeting minutes were also regularly posted to the Teams channel. Following the voting, the committee discussed the results and

provided context as to why they voted a particular way and offered suggestions to improve scores on each of the items.

Rubric	
Theme	Description
Public Benefit	Public benefit and/or user needs are clearly defined and understood
Diverse expertise	The project team is diverse and there is sufficient expert input
Comply with the law	There is clarity on legal requirements for the project
Data limitations	Data for the project are of good quality, suitable representativeness, and reliable
Model limitations	The model is reproducible and likely to produce valid outputs.
Wider implications	There are long-term, continuous evaluation and maintenance structures in place

Table 2: Actions to Assess Project

4. Results of Assessment

Results from the poll are show in Table 3. The committee felt more work needed to be done to improve transparency, accountability, and fairness. Of these three metrics, the committee felt we were furthest along on fairness, but the group needed to make improvements in the training materials provided to advisors and communicate more clearly about how to use and interpret the results they will see for a particular student. In addition, the committee emphasized that this tool is only one piece of student success, and that additional services and support structures need to be in place to help students. The committee also recommended more faculty involvement in the process, as they play a pivotal role in the success of students.

With respect to transparency and accountability, the committee recommended that we develop an internal-facing website, both to inform

students of the use of the model, what kinds of data are included, and how the information is used, as well as to raise awareness among employees. Part of the motivation for more employee awareness is to help develop a culture of data governance and record keeping. For example, metrics on student participation in registered student organizations is sometimes inconsistent and haphazard. Improving these sources of information could lead to additional early indicators of students who may be struggling. The committee also recognized that the model deployment was still in the early stages, and that many of the suggestions will take time to implement.

Accountability Results					
Theme	SA	A	N	D	SD
Transparency	0%	0%	29%	57%	14%
Accountability	0%	14%	43%	14%	29%
Fairness	0%	43%	43%	14%	0%

Table 3: Results for Accountability Assessment

As with the accountability assessment, when asked to assess actions associated with implementation, the committee emphasized that we must do more work to clarify our goals and evaluate our project, (Table 4). To better define the public benefit, we need to improve documentation and awareness (again, with the development of an internal website as mentioned above), and to improve other student services. The committee also recommended more frequent feedback from within the advising community, as well as a continuous monitoring process that is well articulated.

One specific recommendation was to add someone with legal expertise. The institution had recently hired a privacy officer, and this individual was later invited to join the committee. We also considered adding an attorney, especially given that the state in which the institution was located was considering AI transparency and privacy legislation. While the legislation had not passed, one benefit from this committee's work was that it was designed to meet the requirements of the proposed bill, which included a clause on public disclosure and transparency.

Action Results					
Theme	SA	A	N	D	SD
Public Benefit	14%	29%	14%	29%	14%
Diverse expertise	57%	14%	29%	0%	0%
Comply with the law	0%	17%	67%	17%	0%
Data limitations	0%	43%	57%	0%	0%
Model limitations	0%	71%	29%	0%	0%
Wider implications	0%	14%	57%	29%	0%

Table 4: Results for Action Assessment

5. Discussion

Since our original model and committee oversight was put into place, the use of AI in higher education has continued to explode. The proliferation of these tools outpaces not only institutions' ability to understand and evaluate the tools, but also the state and Federal governments' ability to regulate or create policies to protect privacy and ensure fairness in a legal framework, much less within a larger social framework. Some authors have argued that that before any technology is implemented, questions should be asked about whether a non-technical solution would be better and that stakeholders and the public should push back against technological solutions in higher education (O'Neil 2017; Bird et al. 2022). Others have taken a more nuanced perspective, agreeing that a clear problem statement and justification should be articulated, with appropriate oversight and boundaries for an ethical use of the tool (May/Iksal/Usener 2016; Mathies 2018; D Florea/S Florea 2020; Martens 2022l Broussard 2023; O'Donoghue 2023; Abassi 2025). Our institution has implemented policies and practices that align more closely with the latter perspective. Scalable, effective tools often require the compilation and distillation of large amounts of data. Failing to use these data in support of student retention and graduation outcomes often results in solutions that are ad hoc, anecdotal, and simply not scalable given the number of students involved and the typical size of advising student load. However, we agree

that issues of privacy and model oversight are of paramount importance. Ethical use of these models in educational settings requires transparency, explanations of limitations, and clear acceptable use guidelines.

The work of refining and implementing the student support model at our institution is just beginning, as is the work of this ethics oversight committee. One frustration of advisors with the initial model was that, despite high prediction accuracy, it was difficult to understand why a student might be flagged as being at risk for dropping out. While factors influencing the prediction score were shown to advisors, many of these had limited ability to be acted upon (for example, characteristics of the high school). The model has been significantly reworked, and the new version is currently in beta testing. Rather than predict retention in the first year, the revised model changed the objective function and now predicts whether or not a student is at risk for failing a class in the first term. This outcome has advantages in that it is proximal (near-term), it is well suited for prediction given the heavy reliance on administrative data, and factors impacting course grades are often (not always) more actionable from a faculty or advisor standpoint.

Another key revision is the inclusion of a new variable, one that assesses the rigor of the student's coursework for that term and compares it the student's prior academic history. For example, a student with a relatively low entering GPA and a relatively difficult first semester course work (based on historical student performance in a similar set of courses) will be a flag for an advisor that the student may need additional tutoring or wrap-around services. These changes were made with the goal of having more actionable insights from the tool. We recognize that students may withdraw for many reasons, not all of which are academic. The revised algorithm seeks to identify risk factors for failing a class, which is strongly associated for withdrawal due to academic reasons.

In responding to the recommendations for the committee, we are also in the process of examining the LMS data and understanding patterns of log-ins across programs and campuses, to determine the best way to incorporate it into the model. At the current time, the LMS system is used within 97% of undergraduate course sections; this high level of penetration is promising. We

expect that inclusion of these data will help find patterns where students' level of engagement changes or are less engaged than peers enrolled in similar coursework. By focusing heavily on academic and administrative data tied to course performance, advisors will have a better idea of what type of outreach is warranted when a student is flagged as being at risk.

One recommendation that we have yet to implement but has received the most discussion is the inclusion of the student voice (Corrin 2021; O'Donoghue 2023). Many authors have made the case that student awareness of how their data is used and granting students the opportunity to dialogue with faculty and staff regarding information or decisions that may be obtained from ML/AI tools would go a long way to addressing issues of agency, informed consent, privacy concerns, and respect for persons (May/Iksal/Usener 2016; D Florea/S Florea 2020, Corrin 2021; O'Donoghue 2023). The committee strongly recommended for a website and public disclosure, which may eventually be legally required should the state pass legislation to that effect. Additionally, there was strong support for either adding a student to the ethics committee or having focus groups or work groups of students to provide feedback on the model's use.

More debate existed within the committee and advising community around whether to make the score from the model visible to students, either as part of their advising visits or as something they can access on their own. On one hand, the ability to view that information might lead to a helpful discussion with the advisor about course load difficulty or deployment of student support structures such as tutoring or accommodations. On the other hand, we want to make sure that students do not assume blame or self-doubt, creating de-motivation (Lawson et al. 2016). Balancing those two considerations is tricky. It is likely in the coming months as the new model is deployed, we may conduct a pilot test where some students are provided more awareness of how their information is being used in support of retention and graduation efforts at the institution.

This work does have limitations. An early warning indicator is only one piece of student success. This model, especially in its current version, is most useful for assisting in academic success. For example, the information about a

particular student might be useful for having conversations about overall course load difficultly or whether additional academic supports might be useful. While grades and course engagement can be a proxy for other issues, such as mental or physical health concerns, the model would not be able to distinguish among those possibilities. Clear training with advisors about how to use—and not use—the model's results are important. We recognize that language matters, and how we describe this algorithm and its use case is important. The burden of retention does not fall solely on the shoulders of students; faculty and staff also have a role to meet students where they are and help provide the tools they need to be successful.

Secondly, knowing that students are at risk of failing a class is insufficient on its own if institutions do not also establish protocols for responding and assisting students with whatever support they may need, whether it is academic, financial, health related, or something else. Our institution currently lacks a clear way to effectively triage students beyond the faculty/advisor level. Furthermore, this academic warning system needs to be tied more closely to other support systems, such as that used by Student Affairs for students struggling with non-academic issues, such as mental health or food insecurity. More work needs to be done in this space, and we need to not only be able to quickly identify students who may benefit from additional services, but also activate and deploy those resources. Use of these models should not exist in a vacuum, and it is important that institutions build the infrastructure to support students throughout their academic career. To that end, another current focus of the institution is to utilize the CRM more effectively to mobilize services. We hope to also have data to feed back into the model, so if a faculty member or advisor reaches out to a student or directs them to services, we can include that information in the model to update predictions and assess efficacy of interventions.

While work is ongoing, the establishment of this oversight committee, and more importantly, the exercise of the formal scoring the project on a series of metrics to assess its use, has triggered thoughtful and respectful conversations. The committee co-chairs were surprised that generally the committee was less worried about the technical aspects (such as accuracy) or the types of data included in the model, and more concerned with

transparency surrounding the model's use, the limitations of the model, and the interpretation and use of the results by humans. In this sense, the committee has a framework similar to Gillespie (2013), who argues that humans have a larger role in "automated" algorithms than they often like to admit. Many of the committee's suggestions came back to increased documentation, training, and resources for deployment. The suggestions are both specific and achievable, and we will work to implement those going forward. We also want to emphasize that this work should be seen as iterative, ongoing, and collaborative, rather than fixed and determinative, a as key aspects of cyber-ethical leadership (Abassi 2025).

Applying a scoring rubric to a model developed in house requires some humility, (especially in the early stages of deployment), as developers and promoters may find out they need to make significant improvements to either the model or its implementation. In this paper, we have presented a framework that institutions could adopt to fit their specific use case and institutional characteristics. Ultimately, we hope this process and mechanism for continual oversight and review will be useful for institutions who seek to balance the use of powerful ML/AI tools for student success or institutional efficiency with respect for persons, fairness, and privacy. Tools of this type require continuous monitoring and re-assessment to ensure that they continue to work as intended and that both students and employees understand its value.

6. Bibliography

Abassi, Ryma. 2025. "Cyber-Ethical Leadership in Higher Education: A Practice Based Framework for the Digital Age." Journal of Ethics in Higher Education, no. 7.2, 79–101.Bain, Lee J, and Max Engelhardt. 1992. Introduction to probability and mathematical statistics. Vol. 4. Duxbury Press Belmont, CA.

Bauman, Dan. 2022. "Congress should scrutinize higher ed's use of predictive analytics, watchdog says." The Chronicle of Higher Education (June). https://www.chronicle.com/article/congress-

should-scrutinize-higher-eds-use-of-predictive-analytics-watchdog-says.

Benjamin, Ruha. 2019. Race after technology: Abolitionist tools for the new Jim code. John Wiley & Sons.

Biddix, J Patrick, Kaitlin I Singer, and Emilie Aslinger. 2018. "First-year retention and National Panhellenic Conference sorority membership: A multi-institutional study." Journal of College Student Retention: Research, Theory & Practice 20 (2): 236–252.

Bird, Kelli A., Benjamin L. Castleman, Yifeng Song, and Renzhe Yu. 2022. Is big data better? LMS data and predictive analytic performance in postsecondary education. 647. Annenberg Institute at Brown University, September. http://www.edworkingpapers.com/ai22-647.

Birhane, Abeba. 2021. "Algorithmic injustice: a relational ethics approach." Patterns 2 (2).

Broussard, Meredith. 2023. More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech. MIT Press.

Brown, Michael, and Carrie Klein. 2020. "Whose data? Which rights? Whose power? A policy discourse analysis of student privacy policy documents." The Journal of Higher Education 91 (7): 1149–1178. Page 15

Burd, Stephen J. 2024. Lifting the Veil on Enrollment Management: How a Powerful Industry is Limiting Social Mobility in American Higher Education. Harvard Education Press.

Cardona, Tatiana, Elizabeth A. Cudney, Roger Hoerl, and Jennifer Snyder. 2023. "Data mining and machine learning retention models in higher education." Journal of College Student Retention: Research, Theory & Practice 25 (1): 51–75. https://doi.org/10.1177/1521025120964920.

Central Digital & Data Office. 2020. Data ethics framework guidance. September. https://www.chronicle.com/article/congress-should-scrutinize-higher-eds-use-of-predictive-analytics-watchdog-says.

Corrin, Linda. 2021. "Shifting to digital: a policy perspective on 'Student perceptions of privacy principles for learning analytics' (Ifenthaler & Schumacher 2016)." Educational Technology Research and Development 69 (1): 353–356.

DeBard, Robert, and Casey Sacks. 2011. "Greek membership: The relationship with first-year academic performance." Journal of College Student Retention: Research, Theory & Practice 13 (1): 109–126.

Drew, Cat. 2018. "Design for data ethics: using service design approaches to operationalize ethical principles on four projects." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376 (2128): 20170353.

Ekowo, Manuela, and Iris Palmer. 2016. "The Promise and Peril of Predictive Analytics in Higher Education: A Landscape Analysis." New America.

--------. 2017. "Predictive analytics in higher education: Five guiding principles for ethical use." New America.

Florea, Diana, and Silvia Florea. 2020. "Big data and the ethical implications of data privacy in higher education research." Sustainability 12 (20): 8744.

Fountain, Jane E. 2022. "The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms." Government Information Quarterly 39 (2): 101645.

Giffen, Benjamin van, Dennis Herhausen, and Tobias Fahse. 2022. "Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods." Journal of Business Research 144:93–106.

Gillespie, Tarleton. 2013. “The Relevance of Algorithms.” In Media Technologies: Essays on Communication, Materiality, and Society, edited by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot, 167–193.

Government Accountability Office [GAO]. 2022. Congress should consider enhancing protections around scores used to rank consumers. Page 16

Green, Ben. 2022. “The flaws of policies requiring human oversight of government algorithms.” Computer Law & Security Review 45:105681.

Jones, Meg Leta. 2017. “The right to a human in the loop: Political constructions of computer automation and personhood.” Social Studies of Science 47 (2):216–239.

Kimbell, Lucy. 2011. “Designing for service as one way of designing services.” International journal of design 5 (2).

Komljenovic, Janja, Sam Sellar, and Kean Birch. 2025. “Turning universities into data-driven organisations: Seven dimensions of change.” Higher Education 89 (5): 1369–1386.

Lawson, Celeste, Colin Beer, Dolene Rossi, Teresa Moore, and Julie Fleming. 2016. “Identification of ‘at risk’ students using learning analytics: the ethical dilemmas of intervention strategies in a higher education institution.” Educational Technology Research and Development 64 (5): 957–968.

Martens, David. 2022. Data science ethics: Concepts, techniques, and cautionary tales. Oxford University Press.

Mathies, Charles. 2018. “The ethical use of data.” New Directions for Institutional Research 2018 (178): 85–97.

May, Madeth, S´ebastien Iksal, and Claus A Usener. 2016. “The side effect of learning analytics: An empirical study on e-learning technologies

and user privacy." In International conference on computer supported education, 279–295. Springer.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. "A survey on bias and fairness in machine learning." ACM computing surveys (CSUR) 54 (6): 1–35.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. "The ethics of algorithms: Mapping the debate." Big Data & Society 3 (2): 2053951716679679.

O'Neil, Cathy. 2017. Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

O'Donoghue, Kevin. 2023. "Learning analytics within higher education: Autonomy, beneficence and non-maleficence." Journal of Academic Ethics 21 (1):125–137. Page 17

Olhede, Sofia C, and Patrick J Wolfe. 2018. "The growing ubiquity of algorithms in society: implications, impacts and innovations." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376 (2128): 20170364.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. "Social data: Biases, methodological pitfalls, and ethical boundaries." Frontiers in big data 2:13.

Parliment, European, and the Council the European Union. 2016. Regulation 2016/679. Directorate-General for Justice and Consumers, Eurostat. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng.

Reidenberg, Joel, N Cameron Russell, Jordan Kovnot, Thomas B Norton, Ryan Cloutier, and Daniela Alvarado. 2013. "Privacy and cloud computing in public schools."

Reidenberg, Joel R, and Florian Schaub. 2018. "Achieving big data privacy in education." Theory and Research in Education 16 (3): 263–279.

Rubel, Alan, Clinton Castro, and Adam Pham. 2020. "Algorithms, agency, and respect for persons." Social theory and practice, 547–572.

Schonberger, Daniel. 2019. "Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications." International Journal of Law and Information Technology 27 (2): 171–203.

Schwartz, Reva, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, Patrick Hall, et al. 2022. "Towards a standard for identifying and managing bias in artificial intelligence." NIST Special Publication 1270:1–77.

Selbst, Andrew D, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. "Fairness and abstraction in sociotechnical systems." In Proceedings of the conference on fairness, accountability, and transparency, 59–68.

Weinberg, Lindsay. 2024. Smart University: Student Surveillance in the Digital Age. JHU Press.

7. Acknowledgements

The authors would like to thank members of the Ethics Committee for their insight and suggestions. Additionally, the authors would like to thank Nathan Lindsted and Tyler Biggs for their work on model development.

8. Short biography

Stephanie L. Kane is the Assistant Vice Provost of Institutional Research and a doctoral candidate in the Individual Interdisciplinary Program at Washington State University in Pullman, WA. slkane@wsu.edu

William B. Davis is the Vice-Provost for Academic Engagement and Student Achievement and the University Accreditation Liaison Officer at Washington State University in Pullman, WA. wbdavis@wsu.edu