OPTIMIZING SCHOOLS

CASE STUDY: 3

The development of artificial intelligence (AI) systems and their deployment in society gives rise to ethical dilemmas and hard questions. This is one of a set of fictional case studies that are designed to elucidate and prompt discussion about issues in the intersection of AI and Ethics. As educational materials, the case studies were developed out of an interdisciplinary workshop series at Princeton University that began in 2017-18. They are the product of a research collaboration between the University Center for Human Values (UCHV) and the Center for Information Technology Policy (CITP) at Princeton.

For more information, see **http://www.aiethics.princeton.edu**

DIALOGUES ON
AI AND ETHICS

In 2012, Minerva High School, a public school in Pittsburgh, PA, with nearly 3,000 students and 180 classroom teachers, reached a depressing milestone. That year, the student dropout rate hit its highest point ever at nine percent. This meant that nearly one out of every ten students who entered the school as a freshman left without graduating at all. And only 55 percent of Minerva students were able to graduate on time, compared to the state average of 76.4 percent. The school board responded to these dismal statistics by demanding that the school principal, Mr. Vulcani, address this growing problem, or risk the loss of funding – even threatening a possible shutdown. Mr. Vulcani feared for the school's future, but he was at a loss. Students were habitually disengaged, and teachers had been demotivated by frequently shifting incentives, the imposition of new pedagogical approaches and increasingly negative assessments.

Mr. Vulcani met with school board members who suggested he put the vast and varied datasets the school had already collected about its students' behavior to use. In addition to the typical performance, disciplinary and attendance records, the school had given students scannable ID cards that registered when they were used to open the library doors, confirm attendance in class, purchase snacks or lunches and the like. Even the school's Wi-Fi network tracked students' internet use and monitored their movements throughout the campus with an impressive degree of accuracy, using their mobile phones as proxies. These measures produced large quantities of data. The school board members suggested that developments in data science and machine learning could be applied to this information in order to shed light on the causes of what appeared to be an irreversible trend towards high dropout rates. Understanding what causes students to drop out might suggest appropriate interventions and could inform the creation of new incentive structures for teachers and students.

Mr. Vulcani took these suggestions to heart and contracted a local data science company, Hephaestats, that promised insights into business processes through novel approaches using artificial intelligence. Together they formulated these specific goals:

1. To identify predictors of student disengagement as an indicator for dropping out, and to apply machine learning tools to these predictors in order to flag at-risk students;
2. To equip teachers with fine-grained information to allocate resources and assist at-risk students by suggesting specific interventions, such as talking to the student directly, adjusting their workload and schedule, contacting parents or meeting with the student's counselor;
3. To be transparent in the way the system is used.

Mr. Vulcani and the school board agreed to provide Hephaestats with their existing databases, spanning several years, and gave them access to new data as it was collected. Students and parents were not notified of this agreement, nor were they given the opportunity to opt out. Knowing from previous experience how difficult it is to get parents and administrators to come to a consensus on any new initiative, and given the urgency of the situation, Mr. Vulcani believed this was for the best. Besides, he argued, this decision was supported by the school board and fell within his general mandate to promote positive educational outcomes for all.

Discussion Question #1:

How should decisions to adopt AI technologies be made? In this case, it came from above – a suggestion from the school board, implemented by the principal. Who are the other relevant stakeholders? Should they have been involved in the decision to utilize Hephaestats? To what extent? How does the decision not to include an opt-out option affect the legitimacy of the school's actions?

Upon receipt of the student data, Hephaestats began with a broad policy of data analysis looking at a large number of predictors, ranging from various student demographics (e.g. race, ethnicity, gender, mobility, address, home life) to academic factors (e.g. grades, GPA, test results, history of disciplinary action, attendance) to teacher statistics (e.g. certifications, degrees, percent of students failing per class, years of teaching). Hephaestats then harvested new data for a full academic year, allowing its machines to correlate the data of students that previously dropped out with information about current students in order to recognize patterns. Ultimately, Hephaestats was able to produce its own synthetic data by generating inferences that would not have been possible without inputting the school's original data into its algorithms.

Discussion Question #2:

Did the school violate the privacy of its students by sharing their data with Hephaestats? If so, is this breach justifiable, and on what grounds? Would you feel differently if you were the parent of an at-risk student versus the parent of a valedictorian?

Using all this data, Hephaestats was able to identify eight key indicators that, in combination, predicted whether a student would drop out with 92 percent accuracy. The reasons ranged from predictable administrative issues (e.g. an overly ambitious or unsuitable combination of chosen courses, course scheduling) to external factors (e.g. balancing a job with school, domestic responsibilities), as well as previously unconsidered factors (e.g. poor nutritional options in the school cafeteria). Hephaestats then supplied teachers with profiles of at-risk students that helped them better understand why an individual student might be struggling and suggested targeted approaches for helping him or her. These treatment protocols included things like tutoring, modifying assignments, talking with the student's guardians and, in certain extreme cases, contacting local authorities to address possible problems at home.

Some teachers readily followed the recommendations made by Hephaestats, and there was an immediate boost in student engagement. At the same time, the administration used the information provided by Hephaestats to adjust certain aspects of the campus environment in order to nudge students towards better behavior. For example, students were encouraged to change some of their courses to make their schedules more manageable. Some of the most popular sugary foods were also removed from the cafeteria, and all teachers were instructed to provide more attention to struggling students, rather than the students who were already expected to do well.

By the end of the 2016-17 academic year, Minerva High School appeared to have made an impressive turnaround. The four-year graduation rate had risen from 55 percent to 85 percent – a praiseworthy increase, according to the superintendent, that was distinctly higher than the district's average. The dropout rate similarly improved, from nine percent to only five percent. Nearly all Minerva students now left school with their diplomas, and a higher percentage than ever before were being accepted to four-year colleges. In an interview with a local news reporter, Mr. Vulcani praised the work done by Hephaestats, which he argued was instrumental to this improvement in student outcomes.

Discussion Question #3:

How might we define a successful outcome for Minerva High School? What was the school hoping to optimize for by contracting Hephaestats? To what extent is this end legitimate? In order to achieve this end, did Hephaestats ask the right questions and use the right proxies?

Discussion Question #4:

Graduation statistics did, indeed, improve after Hephaestats came on the scene. But correlation doesn't necessarily imply causation. What are some alternative explanations for the improvement in dropout rates?

The seeming success of Hephaestats' approach was overshadowed to some extent by concerns raised by students and their parents when they were finally told about Hephaestats' involvement. Not only did they learn that Hephaestats had been using student data to make its recommendations, but they were informed that the school's data collection and use policies were going to continue into the foreseeable future in order to maintain the system's accuracy. Effective as Hephaestats had been in improving educational outcomes, this decision, delivered from above, made many students and their parents uncomfortable. Emboldened by their critiques, some teachers and administrators also began to publicly voice criticisms of and opposition to the new system.

Ethical Objection #1: Privacy

Critics claimed that the data provided by the school to train Hephaestats' algorithms amounted to a fishing expedition, whereby vast amounts of data were provided without regard for its sensitivity. Surely, they argued, the school was breaching privacy laws or at least some expected norms by handing over such a broad range of data to a commercial entity. And even if the goals were laudable, it didn't seem right to collect and use student data without first getting consent from students or parents. If the community allowed such a blatant violation of privacy in this instance, they worried it would set a bad precedent.

Ethical Objection #2: Dehumanization of Students and Faculty

Many students didn't like the idea of being treated as research subjects – even if it was for their own good. Several reported feeling like subjects to be optimized in a living lab, where every action could be scrutinized by an invisible, all-powerful computer. Critics acknowledged that the new AI approach to enhancing educational outcomes might be beneficial for those students who could be targeted for intervention, but that overall, the system seemed to be primarily geared towards improving the school's ranking. This sidelined the school's subtler aims, such as providing a safe space for students to fail, the opportunity to experiment with interests and ideas, and a nurturing environment in which students could grow to be thoughtful, self-sufficient adults. Teachers argued vehemently that their training, experience and intuitions were being overridden by the new AI system.

Ethical Objection #3: Transparency

In general, students, parents and teachers felt they had been forced to trust a process for evaluating risk and identifying solutions that they could not scrutinize themselves. If they were to be beholden to such an authoritative, powerful and opaque system, then they wished to be involved in the design process as participants, rather than serving merely as its data subjects. They also insisted on a procedure for appealing disputed results and requested the source code of Hephaestats' systems. To emphasize the urgency of their request, they threatened to file a formal Freedom of Information Act (FOIA) request.

In addition to these ethical considerations, the school was criticized for over-enthusiasm about using artificial intelligence as a way to modernize education. Several teachers argued that the machines were just telling administrators the same things teachers had been saying for years. Working on the frontlines, teachers often know why individual students are failing and what to do about it. They just rarely have sufficient institutional support to make the necessary changes. Thus, these teachers attributed much of the improvement in graduation statistics after the introduction of Hephaestats merely to the "Consultant Effect."

One of the school's math teachers had a slightly different take, but similarly questioned the "wisdom" of AI systems like Hephaestats. She wrote an op-ed in the local newspaper, describing her concern that the new system was simply rehashing statistical methods, rather than using a new form of intelligence. The teacher explained that all the flaws, limits and biases that are well-known in statistics were being swept under the rug

by rebranding the system as artificial intelligence, thereby "blackboxing" the processes. She argued that blind faith in such an unchecked system of statistical governance may lead to the kinds of long-term problems that still plague the field of statistics.

Discussion Question #5:

The rhetorical decision to call a technology "AI" imbues it with a certain mystique. But quantitative and statistical methods, such as those used in many AI systems (including Hephaestats) inherently involve generalizations. While there is value in using statistics to understand social problems and make predictions, the methodologies may not be useful on an individual basis. What is the danger of calling a system that deals in particulars "AI"? What are the advantages?

Representatives from Minerva High School and Hephaestats met with concerned students, parents and teachers to respond to their worries about the new system. While the school admitted to handing over vast amounts of data, they considered the privacy concerns unfounded. First, the databases were all pseudonymized and no record was being kept of the link between data points and students' identities in the raw dataset. Second, even if the data had been identifiable, they argued that they did not need to seek consent from students or their parents because the data was collected for the legitimate purpose of improving educational outcomes.

For its part, Hephaestats resisted calls to release its proprietary algorithms. Representatives from the company argued that the source code would be meaningless to the public. The original code was developed as a means to enable Hephaestats' algorithms to learn from patterns and correlations in the existing data, but the code had long since evolved into complex neural networks with millions or even billions of nodes. It would be impossible for Hephaestats (or any other machine learning experts) to adequately interpret the current functioning of such a system. However, Hephaestats did provide some tools that would allow students to understand which data points were used to recommend a particular course of action. Given the state of the art of such technology, this was the best they could feasibly do.

Hephaestats' representatives agreed with the math teacher that their system was largely based on statistical methods. However, they noted that the system makes use of some recent developments in the field of machine learning, which is both a subfield of artificial intelligence and predictive statistics. Since the term "artificial intelligence" is more commonly used, they decided to use that labeling, while making sure to correct their processes and calculations for the known limitations in the field. They did not intend to mislead.

Reflection & Discussion Questions

Privacy: When designing a system of AI governance, some trade-offs are inevitable. For example, individual privacy considerations often must be balanced against the desire to achieve legitimate social ends. The extent to which specific values are embedded in systems reflects the priorities and preferences of the systems' designers. And the extent to which users accept and utilize these systems is likewise reflects users' priorities. In the Minerva case, the school board and Mr. Vulcani decided that some infringement of students' privacy was a reasonable price to pay for a lower dropout rate.

- How should decisions about the appropriate balance between privacy and improving educational outcomes be made? Who should be involved in this process and to what extent?

- How does the issue of privacy change in the school setting? Would the appropriate balance between privacy and social welfare be different in a private school? In an office setting? Think of this question using both ethical and constitutional law frameworks.

Autonomy: Autonomy is an individual's ability to make decisions for herself and act upon them. Hephaestats, like many other AI systems, may compromise this value. For teachers, the use of Hephaestats' system could mean that their priorities and judgements are overridden. The issue of autonomy is still more complex for students being targeted for interventions. Students who are still minors are not thought to have the same degree of autonomy as adults, either in theory or law. Treating minors as less autonomous can be good in terms of protecting more vulnerable members of society, but also raises some difficult questions about paternalism.

- Should Hephaestats provide students with their risk profiles? Should students have a right of appeal? Should they be able to opt out of being assessed? Would it be possible to include them in decisions regarding the design and deployment of Hephaestats, and if so, how?

- Hephaestats offered several options to address the student dropout rate at Minerva High School, but they mostly emphasized a "nudging" model. This meant that students typically did not receive explicit mandates or directives; rather, they were nudged along in certain directions through changes to the incentive structures (e.g. by making sugary foods less available). Nudging represents a softer form of influence, but it is influence nonetheless. AI systems designed to "nudge" are justified on the basis that they adjust choice architecture to help people make the right decisions. But the right decisions for whom? Who decides? Are nudges a better way of producing desired outcomes than more explicit exertions of power? Why or why not?

Consequentialism: Some people argue that certain actions are impermissible regardless of what good outcomes they might bring about; others believe that the ends may justify the means. The Minerva administrators and their partner, Hephaestats, had both good ends and, they argued, appropriate means. But in complex AI systems, it may be quite challenging to even keep track of the various means in use. If all these means must be evaluated independently of the ends they're used to bring about, it may be very difficult to evaluate the permissibility of different actions. Furthermore, when AI is deployed to solve real world problems, each step of the implementation must be tracked as well. Considering the difficulty of assessing each of these steps in their entirety, school officials and Hephaestats preferred to focus on their noble end of reducing the student dropout rate.

- Even if nearly everyone felt the school dropout rate was a problem, not all stakeholders agreed with Hephaestats about the appropriateness of their means, namely, their use of student data without consent to produce un-auditable results. These dissenters might argue that the way Hephaestats went about reducing the dropout rate undermined its ultimate success in achieving this "noble" end. What would you say?

- If we accept that all significant stakeholders ought to have a voice in determining the values they want their communities to promote, does it follow that they should be involved in decision-making about the means of achieving those ends as well? How would schools go about including them?

Rhetoric: The use of language is very important, especially in framing and describing new, developing technologies. Hephaestats chose to label and promote their IT solution as "artificial intelligence," but they also could have labeled their approach as a matter of social science or statistics, instead.

- Was Hephaestats right to call its technology "AI"? If not, how should they have formulated their description of the system to the school board, students, parents and teachers?
- What are the implications of calling something "AI"? What kinds of political and social intentions does that decision reflect? Does the "AI" labeling evoke a kind of responsibility beyond what is typically expected from an IT system? If so, what could be done to mitigate these concerns?

AI Ethics Themes:

Privacy
Autonomy
Consequentialism
Rhetoric