The New Atlantis
A Journal of Technology & Society

Machine Grading and Moral Learning

Joshua Schulz

In April 2013, an online education enterprise called edX—a joint project of Harvard and the Massachusetts Institute of Technology—announced that it had created software capable of automatically grading student essays. The news, coupled with the expectation that the software will soon become widely used, generated a heated debate. Speaking to the *New York Times*, edX president Anant Agarwal argued that the software is superior to traditional grading methods because it provides "instant feedback," making it possible for students to rewrite and improve their essays until they get a passing score from the machine. Critics responded that the software relies on seemingly irrelevant or arbitrary criteria—like the presence of particular keywords in the essay—that don't necessarily indicate coherent thought, and may allow the software to be easily duped.

The software is based on machine-learning techniques derived from artificial intelligence research. Using essays already marked by a human grader, along with rubrics that associate quantifiable features of a text—such as the presence or absence of specific terms or phrases—with an evaluative score, the software creates a model for predicting the score a human grader would assign to future essays. The statistical methods behind these learning techniques even allow the software to create "confidence values," which indicate the likelihood that the predicted grade will match that of a human grader. Professors can then re-grade the essays scored by the machine to improve the accuracy of the software.

The software's makers argue that, because its learning model is based on human-generated scores, there is an assurance of both relevance and quality control. The point of the software, in their view, is not to replace professors but to increase their efficiency, especially in large classes. The software is particularly designed for use in the new realm of "massive open online courses" (MOOCs, in the unfortunate acronym), where, like in traditional classrooms, a large class size poses a challenge to learning how to write. This software, in theory, allows graded essays to be a real component of these courses, making the writing process itself more fruitful than in classes where students might otherwise get fully engaged feedback from the professor only on their final draft.

Joshua Schulz *is an assistant professor of philosophy at DeSales University.*

Supporters of edX, like the University of Akron's Mark D. Shermis, point out that many critics of the software come from good colleges, where professors teach smaller-than-average classes and so are able to provide students with above-average feedback on their written work. For everyone else, as Agarwal puts it, "the quality of the [software's] grading is similar to the variation you find from instructor to instructor." Put another way, since most professors grade like robots anyway, the software is arguably a net gain in efficiency. Perhaps the real reason we ought to be wary of the software is that it asks us to embrace this false mindset of what essay grading is.

Functionalism and Its Discontents

The philosophical roots of the software approach to educating the mind lie in a computational view of the mind—a theory of cognition known as functionalism. The theory holds that the mind is defined not by the physical stuff it is made of, but by what it *does*. When we describe aspects of the mind, what matters is the way they transform the mind's *input*, in the form of sense data, into its *output*, in the form of speech and action. The upshot of this theory is that one can have minds without brains, and replicate mental functions using stuff other than brains. In other words, the theory provides a metaphysical foundation for the possibility of artificial intelligence, as well as a loose guide to how to go about creating it.

Consider a classic functionalist analogy. The purpose of a mousetrap is to catch and kill mice. This task can be broken down into smaller tasks: a mousetrap requires a baiter, a trigger, a trapper, and a killer. Many devices made of many materials can satisfy these functional conditions for being a mousetrap. A steel spring on balsa wood or a baited ruler balanced on a bucket of water both catch and kill mice using baiters, triggers, trappers, and killers. A common housecat, though it does not use triggers, baiters, or trappers like a mechanical mousetrap, still fulfills the function of catching and killing mice (if the cat happens to be a good mouser), and so, according to functionalism, even a housecat has a claim to being called a mousetrap, equal to any of these machines.

The functionalist holds that what is true of mousetraps is true of minds. Consider the following scenario (which is a gussied up version of another classic functionalist thought experiment, this one proposed by Alan Turing). Suppose a literature professor with a deep-seated animosity toward mustaches is grading two essays about *Hamlet*. Judging from the essays' mechanics, cogency, and insight into the play, the professor

deems them worthy of the same grade. If it is subsequently revealed that a mustached person wrote one essay and a clean-shaven person the other, it would be plainly irrational for the professor to deny them the same mark: the essays are functionally equivalent, and the fact that the professor has an animus against an irrelevant feature of one of the essay writers does not provide a sound basis for assigning a different grade to one essay over another. By the same token, if it were revealed to a denier of artificial intelligence that a flesh-and-blood human student and an essay-writing computer produced otherwise equivalent essays, it must be an irrational bias to insist that a mind produced one essay but no mind produced the other. To deny this is to succumb to what Ned Block once called neural "chauvinism."

Supported by thought experiments like these, the functionalist theory of mind offers some distinct advantages over competing theories. First, it is open-minded about what kinds of stuff minds are made of. We should not mistake the fact that minds in our world are—so far—made of biological stuff with the claim that mentality must be biologically grounded. Perhaps mentality is "multiply realizable" in neurons, *and* in silicon, *and* in whatever other stuff. Mind isn't an essence; it's a job description. Functionalism is also perfectly consistent with a physicalist understanding of the mind, and gives us the tantalizing prospect of one day forming a completely naturalistic explanation of the formerly mysterious human "soul." With a blueprint of the mind, a map of what neural stuff performs each mental function, we will be able to manipulate and improve the mind, just like a mousetrap.

However, functionalism's critics believe there is a question-begging assumption at its heart. The functionalist argues that if two essays are functionally equivalent, then what produced each essay must be a mind, even if one of the essays was in fact produced by a machine. But as philosopher John Searle famously argued in his Chinese Room thought experiment, the functionalist argument ignores the distinction between *derived* and *original* meaning. Words have *derived intentionality* because we use words as artificial vehicles to express concepts. If the mind is like a wellspring of meaning, words are like cups, shells for transmitting to others what they cannot themselves create. The same is true of all conventional signs. Just as a map is not a navigator and an emoticon is not an emotion, a computer is not a mind: it cannot create meaning, but can only copy it. The difference between a Shakespearean sonnet and the same sequence of letters as the sonnet produced randomly by a thousand typing monkeys—or machines—is that a mind inscribed one with semantic meaning

but not the other. When Shakespeare writes a sonnet, the words convey the thoughts that are in his mind, whereas when a mindless machine generates the same sequence of words, there are simply no thoughts behind those words. In short, functionalism's focus on the behavioral concept of "functional equivalence" forgets that a sign depends on the meaning it signifies. We cannot treat syntactically equivalent texts as evidence of semantically equivalent origins.

Put another way: I don't give plagiarized papers the same grade I give original papers, even if the text of the two papers is *exactly the same*. The reason is that the plagiarized paper is *no sign*: it does not represent the student's thinking. Or we might say that it is a false sign, meaning something other than what it most obviously appears to. If anything, what I can infer from a plagiarized paper is that its author is the functional equivalent of a mirror. As a mirror is sightless—its images are not its own—so too is a plagiarized paper mindless, all of its meaning stolen from a genuine mind. In a Dantean *contrapasso*, I grade plagiarizers with a mark harsher than the F that recognizes an original but failed attempt at thought: I drop them from my course and shake the dust from my feet. We should do the same with functionalism.

John Henry's Retort

Of course, essay-grading software is not functionally equivalent to a professor in the first place, even for the narrow purpose of providing feedback on academic essays. It cannot be, because grading is a morally significant act that computers are incapable of performing. Functionalism, in falsely reducing human acts to mechanical tasks, also reduces the polyvalent language of moral value to a single, inappropriate metric.

If minds are computers, then they should be evaluated by norms appropriate to computers: namely, by their efficiency in mapping inputs to outputs. So if professors and grading software are functional equivalents—outside of the Ivy League, at least—then they should be evaluated using the same criteria: the number of comments they write per paper, their average response time, the degree to which their marks vary from a statistical mean, and so on. This is the latent normative view of functionalism, particularly when it's turned from a philosophical theory into a technical program: if machines can perform some task more efficiently than human beings, then machines are better at it. However, efficiency is not the moral metric we should be concerned with in education, or in other essentially interpersonal, relational areas of human life.

The functional- and efficiency-centric view of technology, and the moral objections to it, have been around for a long time. Look to the tale of John Henry, the steel-driving man of American folklore who raced a tunnel-boring steam engine in a contest of efficiency, beating the machine but dying in the attempt. The moral of the tale is not, of course, that we will always be able to beat our machines in a fair contest. Rather, the contest is a tragic one, highlighting a cultural *hamartia*, namely, the belief that competing with the steam engine on its own terms is anything other than degrading. Consider that, in some versions, John Henry was a freed slave; his freely undertaken labor, in contrast to the pistons of the steam engine, was a sign of economic justice. Consider that John Henry may have been working to support a family, or to save for a homestead of his own, or to buy others out of slavery—that his daily labor was a labor of love. The steam engine worked for no reasons of its own. Consider finally how Henry's voluntary martyrdom demonstrated manly virtues like courage and willpower that the steam engine couldn't replicate even if it had won. For all of these reasons, John Henry's labor possessed a moral significance that the operation of the steam engine did not. His actions were invested with the dignity of the man himself, with what C. S. Lewis called the "weight of glory." The death of John Henry is therefore an objection to the contest itself—a nineteenth-century version of the Turing Test—to its implicit belief that the labor of a man and the operation of a machine are simply equivalent.

This is not, of course, an argument against mechanization as such. There are many tasks where, for the purposes with which we're concerned, mechanizing the task is better than having a person perform it. Digging tunnels and washing dishes are two such tasks that come to mind, even if the mechanization of these tasks leads to a loss of specific virtues associated with them. For example, we might respond to Adam Smith's concern in *The Wealth of Nations* that the division of labor leads to a general loss of citizens' military virtue by pointing out that these virtues are no longer necessary in contemporary society outside of a specialized military force, and that such virtues are available in other areas of life, such as organized sports. Still, as Harvard political philosopher Michael Sandel has asserted about the monetization of civic and familial duties, there are some kinds of mechanization where certain things that we're interested in will be lost and cannot easily be had in another way.

Consider a real and recent case. Yu Suzuki and colleagues at Kyoto Sangyo University have developed "smart" kitchens to teach budding chefs how to cook. One places a fish on the countertop, punches in a recipe, and

the room takes over. After detecting the fish, the kitchen projects dotted lines and animated knives to help the chef fillet the fish, tracks the internal temperature of the fish, orders the chef to flip it when needed, and so on. Here we should ask an obvious question: at the end of a culinary education dispensed in such a manner, what will a good chef know or be able to do? Presumably, the best chefs will be those who most efficiently and accurately carried out the kitchen's directions. These technicians will not need to know anything about food, cooking, or culinary aesthetics at all; this knowledge might even cause the cooks to second-guess the kitchen, becoming a source of inefficiency and error.

It is a familiar functionalist pattern: a tool is invented that imitates the behavior of chefs, and is presumed on that basis to be equivalent to the chef. The result is what the French sociologist Bruno Latour (drawing on Madeleine Akrich) called "prescription"—the normative evaluation and control of human behavior by machines. Suzuki's chefs are going to be judged by the standards appropriate for tools, that is, their efficiency in performing mindless tasks. The traditional virtues or excellences of the chef, including culinary innovation and creativity, will become either irrelevant in the evaluation of the chef or even detrimental to their culinary performance—the functional equivalent of a system error. In short, functionalism does more than metaphysically reduce the mental to the mechanical; it also displaces robust moral and aesthetic categories with narrow technical norms.

Although we could repeat the thought experiment with human soldiers and robotic drones, or any number of other substitutions, our topic is education. So suppose we use edX's software in a smart-classroom to teach professors how to grade essays in the same way smart-kitchens teach chefs to cook. The classroom would project comments and corrections onto student papers for professors to trace, increasing grading efficiency by turning the term paper into the linguistic equivalent of the multiple-choice exam. Would this make professors better professors, or the educational system more educational? Functionalists are committed to thinking that it would—that professors are deficient in ways that can be improved in terms measurable by efficiency, that these terms are appropriate ways of measuring what it is that professors do, and that if professors can be improved in these measures then they should.

The functionalist must think that these are innocent assumptions and meaningless questions, since he believes that whatever acts like a professor is a professor, in the same way that, as far as John Henry's C&O Railroad was concerned, whatever acted like a steel-driving man is a steel-

driver. The functionalist has purged moral qualities from his ontological catalogue. For most everyone else, this approach falsely turns qualitative goods into quantitative, procedural ones—the same mistake made by the rich young man in the Gospels who, having satisfied the Mosaic law, asked Jesus whether anything else was needed for eternal life. It is an error that fails to distinguish between art and technique.

The Moral Art of Grading

Art, unlike technique, requires an understanding of the nature and purpose of one's subject. Being a good builder of *houses* only requires carpentry skills and the ability to follow a blueprint, but a good builder of *homes* understands that homes are for people, and tailors his designs according to their physical, aesthetic, spiritual, and political impact on human beings. For instance, one wouldn't praise an architect for making restrooms out of glass and living rooms out of foot-thick cast iron, since the former would frustrate our modesty and the latter our sociability. Every building presupposes some conception of what a person is, and thus what people are for. The large, lockable bedrooms of modern American homes, for instance, like the "family" rooms designed around television, suggest that what people primarily do at home involves spending time apart until they come together to passively consume entertainment. The architect is a philosopher in brick and mortar. As poems do with words, music with rhythm and melody, and paintings with color and shape, architecture manifests in form, mass, and ornament the aspirations (or perversions) of the human heart. Every art is anthropology.

The same is true of grading. Responding to and evaluating students' written work does more than just describe students, or distinguish them. Grading is also pedagogical: it corrects and informs, rewards and reinforces someone's understanding of the world. Because it has the potential to change a student, grading is a moral hazard. Grading well requires knowing what human beings are for and educating them accordingly; how and why one grades is a confession of one's beliefs about the ultimate destiny of man. A professor is an architect of the intellectual life, making castles of minds and cathedrals of culture—or slums and factories, as the case may be. EdX's software would reduce the professor to the equivalent of a house-builder following a blueprint, oblivious to its moral design. Yet every tool has a design, and therefore an ideology. If we refuse to be passive users of other people's tools—and so tools ourselves—we must ask the designers of such software the same question Plato asked of

government in the *Republic*: what vision of humanity does it presume and therefore seek to bring about? Who guards the guardians from degenerate anthropologies and perverse moral visions?

Consider one such vision. Educators with a corporatocratic or consumerist understanding of their profession—shopkeepers in scholars' clothing—believe we assign grades in order to give employers a quantitative measure by which to compare the skills of potential employees. Like prices, grades signal to potential buyers the quality of the university's products, and thus the potential return on a financial investment. With their belief in the supremacy of self-interest and market forces, the consumerist model of grading gives the merchant university an incentive to inflate student grades when helping their customers (the students), sell themselves to employers. Such grade inflation is contrary to the interests of the capitalists who use grades as signs of the relative quality of potential employees. The job market's demand for informative grades is experienced by professors as administrative pressure to resist grade inflation. But of course, under the consumerist model, keeping marks reasonably low is contrary to the interests of students, who seek to be branded in ways that will attract the highest bidder.

From the students' perspective, their marks do not need to be truthful, or to represent their mastery of any subject. (A student once justified her grade appeal to me not by demonstrating her command of the course material, but by reasoning that it "couldn't hurt, and could help" her graduate and secure a job.) Students who do receive good marks reward their professors with praise on the course evaluations by which universities partially determine which professors to retain or tenure.

In short, the consumerist model of grading makes students, professors, the university, and employers competitors in an unregulated market. It is in the short-term self-interest of each participant to deceive all the others. For instance, it is in the interest of professors to withhold from students details about their grades, and of students to pressure professors into revealing those details in order to negotiate the grade and gain a competitive advantage over other students. Teaching and learning are, needless to say, far from the highest priorities set by the incentives in a consumerist university.

To participate in such a system is to condone its assumptions about the nature of the human person and our theological destiny. In its nominalist and cutthroat universe, grades are seen as results of contingent conjunctions of arbitrary preferences and undeserved power (especially in the liberal arts). This consumerist model of grading and of education

is only appropriate if life is for no more than food and the body for no more than clothing. As a Catholic philosopher, I believe we are made for the bread and wine of a different Eucharist. I believe that wisdom and moral virtues are good for their own sakes, because, as Pope John Paul II writes in his *Gratissimam sane* (*Letter to Families*), human beings, who are the embodiment of these virtues, have been willed from all eternity for *their* own sake, loved into existence in a creative act of infinite generosity. Knowledge comes in several kinds—technical, moral, and scientific—of which the former is instrumentally valuable, and the latter are intrinsically good, because they are *for man.* Our most valuable pursuits are those in which we can have only a disinterested interest—that is, one not governed by purposes outside of these pursuits—such as philosophy, literature, music, and humor, because they are fundamentally forms of love, in Whose image we are made. The soul was made for play and worship, not consumption. As a dying farmer plants for future generations what he will not himself consume, so too is teaching an act of *caritas*, of passing on to another person, for his own sake, the moral, cultural, and intellectual treasure that others entrusted to us. To grade with such a worldview is to exercise something other than the avarice of the consumer model, and something even more than justice.

Religious and secular thinkers alike should be able to grasp this point. Grading should communicate not only what students have achieved but what they can. A professor can encourage intelligent but lazy students with a lower grade than their work strictly merits, and struggling but passionate students with one higher. The principle is far truer for the written responses professors make to student papers. Good professors will challenge a gifted student to address an overlooked problem on a passable term paper purely for the joy of initiating him or her into the life of the mind. They will discourage the well-meaning student from following a line of thought whose path that they know to be littered with intellectual blind alleys and moral dead ends. And what professor hasn't been blessed to discover a new challenge, implication, question, or line of inquiry because of an insightful or prescient student paper?

To grade as if the point were to identify and label mistakes is to grade as mechanics give estimates: this is what is broken and what it will cost to fix. To grade with charity is to treat students not as busted but as becoming. It is to take even their mistaken ideas seriously when they are sincerely offered, by responding with truth and with hope. It means treating grading as a means to continue a conversation older than any of us, and wisdom as both a goal and a common good.

The machine cannot grade with *caritas* any more than it can make mistakes or any more than it can learn. It cannot correct, suggest, encourage, or be surprised. It therefore cannot say whether the executions of Socrates and Christ were tragic or comic or any other weighty thing, nor could it change its mind about such matters mid-program, as an honest professor might do in the face of a novel argument in a cogent student essay. The conclusion that professors and grading machines are functionally equivalent is plausible only if we have already presumed that persons are essentially machine-like, describable and evaluable according to discrete and non-overlapping functions, and that they don't transcend these functions in any way relevant to learning. To use computers to mark student essays is thus to posit a technocratic view of knowledge and an instrumentalist conception of the person—which amounts to denying the existence of persons at all.

A Liberal Education of Love

Let me restate the argument first perspicuously and then suggestively. Grading is an interpersonal process that is constituted by certain norms, attitudes, and virtues that only contribute to human flourishing given a transcendent view of the person. Functionalizing this process displaces these norms with market values and a reductive view of the person, one that is especially antagonistic to the pursuit of wisdom that underlies the liberal arts. Technology that functionalizes an essentially interpersonal process corrodes the moral and eschatological foundations of the liberal arts, trading a pearl of great price for silicon dust. We should not devote ourselves so to the goals of consumerism; we are made for better ends, and a purpose of the liberal arts is to emancipate us from the mundane rather than to further mire us in it. While pedagogical efficiency is a valuable goal, simple technical efficiency is not the way to achieve it, not merely because it's ineffective, but more importantly because it substitutes consumerist ends for higher ones, and so corrupts and undermines the human excellences of wisdom and care that humane education aims to achieve.

EdX's software presents us with a technical solution to a non-technical problem, which is roughly this: students need more people who can provide the detailed and meaningful feedback on their writing necessary to improve their thought and character. For that, recall, is what learning is—a voluntary movement of one's heart and mind toward the truth about the world and oneself, and the ends of each. Since writing is but a sign of our thoughts and desires, one cannot expect to accomplish this learning through the mere syntactic rearrangement of signs.

Learning requires love. All good writers have that teacher, that dear reader in the back of their heads, of whom we ask, "Will he love this?" Only love moves us to imitation, so that we can be good for the beloved rather than merely pleasing. So the student loves the teacher and desires to imitate him, and the teacher loves the student and so imitates—whom? Not the student, for he is the learner. Someone else, then. Can words be copies of copies all the way down—or up? Much as I desire to know the answer to that question, I cannot ask it of a computer, nor do I care to, for it cannot answer, nor care to.

What is required is someone interested in discussing these questions with me. Such friends are in short supply, as perhaps they always have been. Giving us an army of robot imitators will not solve that problem. The underlying motivation for edX's software is the laudable desire to provide what is much in demand but little in supply: patient, loving, discerning educators. But we cannot operationalize this educator as a "feedback provider," any more than we can monetize friendship—for the friend you pay for is not your friend. Attempting to do so will not increase learning efficiency, because no true learning will be taking place—no *metanoia*, no change of heart and mind—but only the rearrangement and redirection of lesser appetites.

So what about Mark Shermis's original argument—that good grading is labor-intensive, forbiddingly expensive, and impossible for professors who teach large classes? The consumerist basis of this argument is now plain. Everything Shermis said is true: it is impossible for professors to grade writing well when they teach large classes. EdX has suggested, and I have denied, that we can solve that problem with technology. I am suggesting, and others will deny, that we could solve the problem by doing away with large universities—or at least with large class sizes—and instead filling small, liberal arts colleges with lots of competent professors who have the time and inclination to learn their students' names.

This disagreement is not primarily about means, as if efficiency or mechanization were intrinsically bad. They're not, or not always. Given that efficiency and relational teaching are not, as edX proponents would have us think, mutually exclusive, the debate is rather about why we should adopt a technology contrary to our ends. The only possible reason could be some other end. It goes by many names. I have called it consumerism. Marx called it, among other things, the commodity fetish. Neil Postman named it technopoly. What we need is a renewed evaluation of the gods or the narratives we take our schools and our culture to serve, since without such a discussion we are merely using evaluative jargon without achieving meaningful thought about the most important ends.