DOI:10.1145/3104986

Jennifer Keating and Illah Nourbakhsh

Viewpoint
Teaching Artificial Intelligence and Humanity

Considering rapidly evolving human-machine interactions.

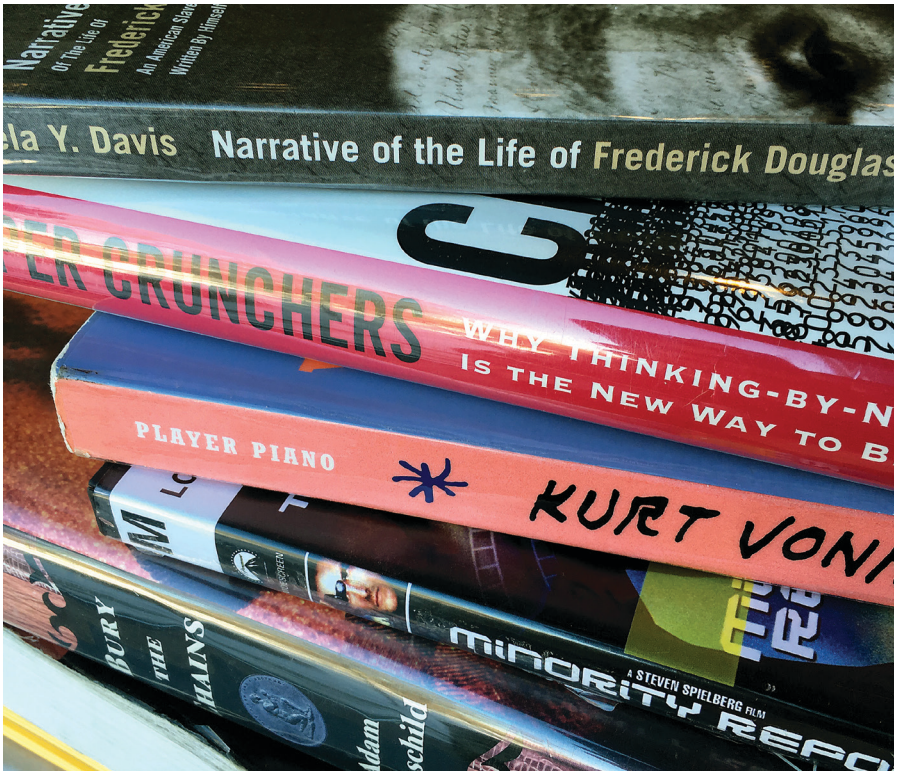

PHOTO BY ALICIA KUBISTA/ANDRIJ BORYS ASSOCIATES

EMERGING ANXIETIES PERTAINING to the rapid advancement and sophistication of artificial intelligence appear to be on a collision course with historic models of human exceptionality and individuality. Yet it is not just objective, technical sophistication in the development of AI that seems to cause this angst. It is also the linguistic treatment of machine "intelligence." Headlines decry the existential threat of machines against humans in various media outlets. But what is really at stake?

Are we truly concerned that we will be surpassed in our capacities as human beings? Or is rhetorical slippage betraying age-old philosophical questions on what it really means to be human? To what degree do our shortcomings in acknowledging human dignity in all populations (regardless of skin pigmentation, linguistic system spoken, geographical location, or socioeconomic position) emerge in questions pertaining to power dynamics between humans and machines? And how might we usefully juxtapose a historic study of our past categorical taxonomies of humanity to more subtly inform our navigation of human-machine relationships? In the fall of 2017 we engaged these questions and more with first-year students at Carnegie Mellon University: 16 students from the School of Computer Science and the Robotics Institute and 16 students from the Dietrich College of Humanities and Social Sciences. In a time of accelerating technological disruption, the next generation of leaders and innovators are ill-equipped to navigate this boundary chapter in human-machine relationships. Perhaps our students can learn from how humans have treated humans to determine viable roadmaps for this challenging moment in our economic, social, and political history, as we mindfully navigate human-machine interactions.

The ways in which machine systems influence our lives have become more explicit in recent years. A chief example that commands popular attention has been IBM's Watson, serving as an informative bellwether for human-machine relations. Its inventors and user community place Wat-

son's clinical knowledge squarely within the social context of the medical community, ascribing agency and a capacity to *learn* to a sophisticated machine with a human name: "Nobody can read it all," Miyano said. "We feel we are a frog in the bottom of the well. Understanding cancer is beyond a human being's ability, but Watson can read, understand and learn. Why not use it?"[6]

Watson's capacity to process data rivals that of a practicing physician and, in some domains, outpaces human abilities. It is positioned as a tool that will rival human capacities in diagnostics to serve as release time for the practicing physician to dedicate more time, energy and intellectual bandwidth to patient-physician interactions. The optimized functions of the Watson apparatus have limitations but they are certainly becoming more sophisticated rapidly: "Before the computer can make real-life clinical recommendations, it must learn to understand and analyze medical information, just as it once learned to ask the right questions on "Jeopardy!" ... The famed cancer institute [Memorial Sloan-Kettering] has signed up to be Watson's tutor, feeding it clinical information extracted from real cases and then teaching it how to make sense of the data. 'The process of pulling out two key facts from a "Jeopardy!" clue is totally different from pulling out all the relevant information, and its relationships, from a medical case ... Sometimes there is conflicting information. People phrase things different ways.'"[2]

Read, Understand, Learn

IBM's Watson is personified as an independent agent in most press coverage. In contrast, at expert conferences like the "Humans, Machines and the Future of Work" conference at Rice University, AI systems like Watson are described as tools. Personification is more tightly regulated when discussed or presented to technologists who are not beholden to the mysteries of the black box, but rather its deconstruction into computational techniques. In the public domain, however, journalists ascribe personhood to the learning machine, which is not necessarily corrected by engineers or physicians, by describing the machine's functions as *reading* and *learning*. Is Watson's information processing and model-building truly reading or understanding? Does such a machine *learn*? Why do we ascribe features historically associated with humanity, subjectivity, and notions of a human self to built machines? And what chapters of human interrelationships are threatened when we readily ascribe human characteristics to engineered systems?

Our society is locked in a stance of both anxiety and ambition in regard to the future of AI. We believe it is crucial that students embarking on undergraduate studies, as budding technologists, writers, policymakers, and a myriad of other future leadership roles, should be better equipped and better practiced in engaging these difficult questions. As automation will be a distinguishing feature in the next chapter of global economies, underemployment threatens the dignity of much of our human labor force. Yet as humans, most individuals would argue they are considerably more than a simple labor force driving a (albeit pervasive and powerful) global economy. An insistence on our capacity to be more than what we might produce as commodities in a market is a distinguishing feature of human dignity in the 21st century. This is a concept, however, that needs to be tested, explored and seriously considered as students prepare to enter this labor force and shape its direction for the coming generations.

Our society is locked in a stance of both anxiety and ambition in regard to the future of AI.

As humans we are readers, we create, we imagine, we strive to understand. Our individual subjectivity allows us to do more than perform specific functions. And yet, present discourse on the potential for AI is oftentimes laced with echoes of our anxieties pertaining to human dignity and its links to work or the distinctiveness of our subjectivity and agency.[4] Will we be 'bested' by the very machines that we build? Will the next generation of technologists be equipped to consider these intended and unintended consequences for the tools they unleash? Or, for now, do we only dream quite wildly beyond technological purviews about the actual sophistication of these tools?

Reading

In the historical context of globalization, labor, human dignity, and education in the West, few rival the narrative potency offered by Frederick Douglass. In *The Narrative of the Life of Frederick Douglass, An American Slave* he writes: "Very soon after I went to live with Mr. and Mrs. Auld, she very kindly commenced to teach me the A,B,C. After I had learned this, she assisted me in learning to spell words of three or four letters. Just at this point of my progress, Mr. Auld found out what was going on, and at once forbade Mrs. Auld to instruct me further, telling her, among other things, that it was unlawful, as well as unsafe, to teach a slave to read."[5]

The power of literacy and its capacity to equip individuals with the necessary tools to dismantle exploitative and unjust systems of power are illustrated in Douglass' work. His capacity to articulate the features of a power negotiation that undermines the very core of the master-slave relationship in a post-Enlightenment era is captured in a human capacity to learn. In the context of the West and its political and social systems, literacy is an opportunity to assert agency. But human-to-human capacities to assert equality, to facilitate Douglass' ability to be 'of no value to his master,' to render him to be 'forever unfit ... to be a slave' due to his capacity to read, are not the tenets used to describe tensions between an AI machine and the tool's "master" [read programmer or

user]. In the context of Douglass' narrative, the prospect of literacy suggests the slave as worth more than his or her labor. Instead, the slave's capacity to learn, to engage in civilized discourse by joining what Benedict Anderson calls "an imagined community," suggests his equal position with other members of the society in contrast to the juridical category of slave as property.[1] It is not the same scenario with AI because the machine is not human, although they are becoming increasingly social. This is relevant because for a long period (arguably a period that we still occupy) humans have treated other humans as tools. The institution of slavery, and its cousin in European colonial systems worldwide, used humans as machines composing a labor force. These were beings who were not extended existential features like agency, subjectivity, individuality, or intelligence. As Joseph Conrad wrote in *Heart of Darkness*, the enslaved Congolese were "black shadows," they were "bundles of acute angles sat with their legs drawn up."[3]

Reading, however, marks agency for Douglass. Reading, in the context of an AI system, suggests anthropomorphic undertones and perhaps humanity. The hierarchical power relationships suggested in these examples are not in the service of hyperbole. The themes they introduce are also not entirely new. The historical analysis of human subjugation offers a rich tapestry regarding agency, identity, autonomy, labor, dignity and citizenry—issues at the heart of how future AI systems and humans will interrelate in our near future. *Reading* and *learning*, the very verbs ascribed today to Watson, are central to Frederick Douglass' discovery of the ways in which illiteracy in enslaved populations reinforces the hegemonic power structure of the master-slave dynamic before abolition. Through reading, Douglass asserts his freedom in deed if not in legal standing.

Understanding

Ascription of features like inconsistency, induction and emotion to machines prematurely suggests essential human characteristics upon our inventions. Yet technologists forge ahead, projecting personhood and agency, coupled with anxiety and uncertainty, upon machines. Even in the case of early light-seeking robots in 1950, W. Grey Walter recognized elements of *human psychology*, from free will to personality: "... the uncertainty, randomness, free will or independence so strikingly absent in most well-designed machines. The fact that only a few richly interconnected elements can provide practically infinite modes of existence suggests that there is no logical or experimental necessity to invoke more than 'number' to account for our subjective conviction of freedom of will and our objective awareness of personality in our fellow men."[7]

Core human characteristics become terms for making sense of complex robotic behavior.

Core human characteristics become terms for making sense of complex robotic behavior, as if complexity is sufficient to justify giving our machines subjectivity. Today, serious legal experts are considering granting personhood to self-driving automobiles, because these AI-driven machines will be so socially integrated into our transportation infrastructure that they need to be individually liable for accidents. Notably, corporations were historically granted limited personhood to shield individual humans from responsibility and blame; personhood ascribed to AI similarly shields both corporations and engineers. Personification trades accountability with convenience for tort and liability, and further with product marketing. Watson, for instance, is named as one technology product across many use cases; but in reality, each disciplinary version of Watson is a separate, custom-made instantiation with its own silo of AI, data store, and interface.

Calendar of Events

February 5–9
WSDM 2018: The 11th ACM International Conference on Web Search and Data Mining,
Marina Del Rey, CA,
Co-Sponsored: ACM/SIG,
Contact: Yi Chang,
Email: garrychang@gmail.com

February 21–24
SIGCSE '18: The 49th ACM Technical Symposium on Computing Science Education,
Baltimore, MD,
Sponsored: ACM/SIG,
Contact: Tiffany Barnes,
Email: tiffany.barnes@gmail.com

February 24–28
CGO '18: 16th Annual IEEE/ACM International Symposium on Code Generation and Optimization,
Vösendorf/Vienna, Austria,
Contact: Jens Knoop,
Email: knoop@complang.tuwien.ac.at

February 24–28
PPoPP '18: 23nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming,
Vienna, Austria,
Co-Sponsored: ACM/SIG,
Contact: Andreas Krall,
Email: andi@complang.tuwien.ac.at

February 25–27
FPGA '18: The 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays,
Monterey, CA,
Sponsored: ACM/SIG,
Contact: Jason Anderson,
Email: janders@eecg.toronto.edu

March

March 5–8
HRI '18: ACM/IEEE International Conference on Human-Robot Interaction,
Chicago, IL,
Contact: Takayuki Kanda,
Email: kanda@atr.jp

Sample of syllabus keywords (left column) and related materials for analysis in seminar (from Williams[8]).

Agency	Frederick Douglass: *Narrative of the Life of Frederick Douglass, an American Slave* *Black Mirror: Men Against Fire*
Self	*Black Mirror: Be Right Back* Jerrold Seigel: *The Idea of the Self*
Technology	Adam Hochschild: *Bury the Chains* Werner Herzog: *Lo and Behold*
Equality and Exploitation	Andrew McAfee: *The Second Machine Age* *Star Trek: The Measure of a Man*
Surveillance	*Minority Report* Ian Ayres: *Super Crunchers*
Labor and Digital Labor	Joseph Conrad: *Heart of Darkness* Simon Head: *Mindless, Why Smarter Machines are Making Dumber Humans*
Citizen	Kurt Vonnegut: *Player Piano* *Black Mirror: Hated in the Nation*
Narrative	Richard Powers: *Plowing the Dark* David Herman: *The Cambridge Companion to Narrative*

Learning

Rapid progress in AI/machine learning and its central role in our social, economic, and political culture signals its salience to the next generation of students entering universities. Building next-generation AI is currently a hot topic. At Carnegie Mellon, we have no trouble filling such classes. And yet, a nuanced understanding of the contributions that technologists are currently making to the world, an indication of how the next generation of computer scientists, engineers, and roboticists might shape the world that humanists and social scientists study, is not at the forefront of our undergraduates' minds. So, how might we ensure this is something they consider throughout their undergraduate career? And that, instead, societal consideration shapes their undergraduate studies from their first year onward? We propose to introduce *AI and Humanity* in the first term of the undergraduate career. Humanities students will sit in class beside their colleagues from the Robotics Institute and the School of Computer Science. They will be taught each class by a team of faculty with an intertwined pedagogical approach: a roboticist and a humanist.

Artificial Intelligence & Humanity is part of a new fleet of first-year courses called Dietrich College Grand Challenge Interdisciplinary Freshman Seminars. These encourage faculty teams to propose courses that attend to historically persistent problems facing humanity, demonstrating an interdisciplinary approach to attending to these problems whose solutions continue to elude us or demonstrate boundary work that a single discipline is often ill-equipped to solve. By harnessing the methodological approaches of various disciplines to demonstrate the complexity and the range of approaches to problem solving in the academy, students are exposed to argumentative structures and efforts to juxtapose historical human-to-human relationships with future narratives of human-to-AI relations.

In Artificial Intelligence & Humanity, students will respond to historical examples of negotiations of power between individuals and communities, then develop language to describe contemporary and historical taxonomies of human-to-human and human-to-machine power relationships. Starting with a survey of narrative forms that explore human relationships that include written memoirs, dystopian television shows, documentary films and science fiction novels, students will consider the various ways in which we narrate our relationships between humans from a variety of perspectives. They will consider how these relationships might manifest in our contemporary consideration of human relationships to machines. We take inspiration from Raymond Williams' *Keywords* and the Key Words Project (http://www.kewords.pitt.edu) to create a conceptual structure of core themes that will guide the semester (see the accompanying table).

In this interdisciplinary course, students will be introduced to both the historical development of AI and to the current state of the art. As we engage with core themes of power negotiations, political implications for advancing technology, and cultural response, students will use terminology from Key Words to build conceptual maps that make sense of technological advances and their societal implications. Students will develop mixed media 'futuring' assignments by semester's end that offer speculations on the future of human relationships to machines. Working in groups, they will create their own narratives, synthesizing a future ethic based on course materials and explorations. While this course is a first experiment in connecting the freshman experience to socio-technical issues relevant to all, we hope that several iterations of course refinement and deployment will yield an approach that can serve as a valuable scaffold for AI and humanity across institutions. C

References

1. Anderson, B. *Imagined Communities*. Verso, London, 1991.
2. Cohn, J. The robot will see you now. *Atlantic*. (Mar. 2013).
3. Conrad, J. *Heart of Darkness*. Bantam Books, NY, 1981, 26–27.
4. Deleuze, G. and Foucault, M. Intellectuals and power: A conversation between Michel Foucault and Gilles Deleuze. *L'Arc* (Mar. 4, 1972).
5. Douglass, F. *Narrative of the Life of Frederick Douglass, an American Slave*. Penguin Books, NY, 1982, 78.
6. Gaudin, S. IBM: In 5 years, Watson A.I. will be behind our every decision. *Computerworld* (Oct. 27, 2016).
7. Grey, W. An imitation of life. *Scientific American* (1950), 42–45.
8. Williams, R. *Keywords: A Vocabulary of Culture and Society*. Oxford University Press, Oxford, 1983.

Jennifer Keating (jkeating@andrew.cmu.edu) is Assistant Dean for Educational Initiatives in Dietrich College, Carnegie Mellon University.

Illah Nourbakhsh (illah@cs.cmu.edu) is Professor of Robotics at The Robotics Institute, Carnegie Mellon University.