Frankenstein's Legacy

Four Conversations about Artificial Intelligence, Machine Learning, and the Modern World

Edited by Brad King

Carnegie Mellon University: ETC Press

Pittsburgh, PA

FRANKENSTEIN'S LEGACY

FRANKENSTEIN'S LEGACY

FOUR CONVERSATIONS ABOUT ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, AND THE MODERN WORLD

Edited by Brad King

Carnegie Mellon University: ETC Press

Pittsburgh, PA

ISBN: 978-1-387-33380-6 (Print)
ISBN: 978-1-387-33420-9 (Digital)

Contents

About This Project

Frankenstein's Legacy: Four Conversations about Artificial Intelligence, Machine Learning, and the Modern World is a collaboration between Carnegie Mellon University's ETC Press, dSHARP, and the Carnegie Mellon University Libraries in collaboration with the Alumni Association's CMUThink program.

This book is part of a university-wide series celebrating the two-hundredth anniversary of the publication of Mary Shelley's *Frankenstein.* This book project specifically sprung from the panel Frankenstein 200: Perils and Potential panel hosted by Digital Scholarship Strategist Rikk Mulligan.

Each of the four panel participants — Jeffrey Bigham, David Danks, Barry Luokkala, and Molly Wright Steenson — sat down with ETC Editor Brad King for wide-ranging discussions about artificial intelligence, machine learning, and the impact of these technologies on the world in which we live. Those conversations were edited into the manuscript that you're now reading.

The book — part of the ETC Press "In Conversation With" series — is a conversational examination and explanation of some of the most powerful technological tools in our society.

Promethean Shadow

The Frankenstein Complex, Creations & Consequences

Rikk Mulligan

Mary Shelley's novel *Frankenstein: or, The Modern Prometheus* was first published two hundred years ago on March 11, 1818. Many critics focus on Victor Frankenstein as a figure of hubris in his pursuit of knowledge without limit, yet more in line with this collection is that in his drive to create life he did not stop to think of the consequences nor what was to become of the result. Shelley's novel warns of the possible sacrifices for knowledge and hints toward the costs to man and society, how new knowledge can redefine human existence and experience. Science and technology have rapidly advanced Shelley's time; sometimes too quickly for creators to fully consider the ramifications of new discoveries. These themes of *Frankenstein* have been reinterpreted and applied to debates regarding atomic weapons, nuclear energy, cloning, bioengineering, robotics, and artificial intelligence (AI). This collection brings together CMU scholars in the Arts, Humanities, and Sciences to consider the relevance of Shelley's novel today, particularly how it helps frame the responsibility of investigators to consider the consequences of artificial intelligence and a technologically-augmented human society.

Victor Frankenstein and his Creature have become global icons thanks to the film adaptations of the past century as the mad scientist and his lumbering, typically mute Creature. However, in the novel, Frankenstein is no doctor and his Creature is an autodidact who teaches himself French and quotes John Milton's *Paradise Lost*. In Shelley's tale, after months of effort and a final burst of manic activity, Victor reanimates the collection of human parts he has stitched together to become his Creature. Having achieved this goal, however, Victor recoils in horror from the product, running from him not one but twice, leaving a newborn being who quickly learns to feel, reason, and seek revenge. Although he started his existence gentle and generous, the Creature is reviled, rejected, and attacked by every human he meets. Driven to hatred and revenge, he stalks his creator, isolating Victor by killing his family, friends, and new bride. Isaac Asimov coined the term "Frankenstein Complex"[1] to refer to this theme of creations that turn on their creators, as the Creature turns on Victor, inspiring calls for ethics in scientific investigation and forethought in creative innovation.

Frankenstein began as a short story in 1816, penned by 18-year old Mary Wollstonecraft Godwin during a competition inspired by a collection of German ghost stories. Although George Gordon, Lord Byron, initially dismissed her story (and failed to complete his own), Mary's then-lover and future husband, Percy Bysshe Shelley, encouraged her to turn it into her first novel. She completed the work in 1817. The novel was anonymously published the following spring, but it was not until the lightly-edited second edition was released in 1823 that Mary Wollstonecraft Shelley was credited as the author.

The fate of Victor Frankenstein is read as a warning of the cost of knowledge, his obsession makes him a Promethean or Faustian figure in his transgression against the laws of gods and nature. The novel's title invokes the myth of Prometheus, the titan who brought the gift of fire, sometimes interpreted as the spark of imagination or creativity, to humanity. This act so angered Zeus, that Prometheus was chained to a rock, where a vulture returns each day to consume his freshly regenerated liver. Victor has also been connected to doctor Faustus, whose quest for forbidden knowledge led to a bargain with the devil and ultimate damnation to hell. Shelley's influences could have included both Christopher Marlowe's 16th century play, *Dr. Faustus*, and Johann Wolfgang von Goethe's *Faust*. Although the influence of the Enlightenment can be seen in the initial childlike innocence of the Creature, he also becomes a lost or damned soul after being radicalized by the books he discovers, in particular John Milton's *Paradise Lost*. The Creature quotes both Adam and Lucifer when he confronts Victor about his creation and abandonment, strengthening the connections between him and these outcasts. While supernatural imagery and references to demons, devils, and the occult are part of the Gothic romance, Shelley also uses them to highlight contemporary natural philosophy–or as we would call it, science.

Frankenstein bridges the medieval to the industrial, beginning Victor's quest for knowledge with medieval theology, and Classical Greek and Roman philosophy, before turning to the natural sciences of the early 19th century. Prior to attending university Victor pored through the works of Albertus Magnus, 13th century theologian, philosopher, and alchemist; the Early

[1] Asimov, Isaac. "Little Lost Robot" *Astounding Science Fiction*; April 1947.

Modern writings of Swiss alchemist and astrologer Paracelsus (Theophrastus von Hohenheim, 1494–1541); and the occult philosophy and medicine of Heinrich Cornelius Agrippa (1486–1535). Paracelsus describes the crafting of artificial life, the "chymicall homunculus" in his treatise *De Rerum Naturae* (1573)[2], and Marlowe's character Faustus proclaims that he will become greater than Agrippa (Act 1, Scene 1, Line 119), both of which point back to Victor's labors and foreshadow his destruction. However, Victor's father not only decries their works as rubbish, but also introduces Victor Benjamin Franklin's experiments with electricity. Once Victor begins his studies at Ingolstadt, the lectures and laboratory work also refers to William Harvey, Robert Boyle, and lectures on natural philosophy, mathematics, and chemistry.

Beyond her use of natural philosophy, Mary Shelley is called the mother of science fiction primarily because she uses the Creature to ask questions of what it is to be human and a part of human society. Inspired by ghost stories and quoting epic poetry, Shelley uses scientific experimentation as plot devices; although she provides details of theory and gestures towards instruments and laboratories, these are not her focus. Waldman's initial talk with Victor nods towards the ancients and authors of earlier natural philosophy as those who set grand goals, yet Shelley uses him to channel Victor's path toward mathematics and chemistry. In his history of science fiction, the *Trillion Year Spree* (1986), Brian Aldiss argues that while the novel is written in the Gothic mode, it is also inspired by scientific investigation in a time before the term "scientist" had been coined. Before he plunges into burial vaults and charnel houses for the parts to construct his Creature, Victor had already excelled in the classroom and laboratory, even developing new technology by improving the instruments used by his professors. Yet these are cosmetic references to experimentation and innovation; they do not make Frankenstein science fiction, rather it is how the Creature troubles the knowledge of the divine and natural law by blurring the line between life and death and asking questions about what it is to be human and a monster or outcast.

Over the past two centuries, Shelley's tale has been adapted to stage, film, television, graphic novels, and videogames. Both the Creature and the creator as "mad doctor" or monomaniacal scientist have become cultural icons recognized around the world. *Frankenstein* was so successfully adapted to the stage that

[2] Campbell, Mary Baine. "Artificial Men: Alchemy, Transubstantiation, and the Homunculus." *Republics of Letters: A Journal for the Study of Knowledge, Politics, and the Arts* 1, no. 2 (April 30, 2010): http://rofl.stanford.edu/node/61.

Shelley's father, William Godwin, edited the novel to capitalize on the story's surge in popularity to publish the second edition. This 1823 stage adaptation focuses much more attention on the laboratory, introducing an assistant (Fritz not Igor). The 1931 Universal Studio film adaptation directed by James Whale drew heavily on the play–it set most of its action in laboratory and castle, and dropped the sweeping pursuits across Europe and into the Arctic, but also left Victor barely clinging to life at the end. In this film Boris Karloff is cast as a heavily made-up Creature, one who does not speak, and so is stripped of his Miltonian shadow but given the now-iconic flattened head, stitches, and neck-bolts. This image has become so recognized that Universal was able to copyright it and has threatened competing studios with legal action since the 1930s. Hammer Studio's *The Curse of Frankenstein* (1957) catapulted that studio into a series of box office successes and nearly twenty years of gothic horror film production. In seeking to innovate, Hammer's version created the link between the Frankenstein Creature and Dracula, and reset Victor's experiments in the late Victorian-era. *Frankenstein* films have proliferated since the 1970s. Many of these adaptations are quickly-forgotten low-budget products, although several notable films including *Young Frankenstein* (1974), *The Bride* (1985), *Mary Shelley's Frankenstein* (1994), *Frankenstein Theory* (2013), *Victor Frankenstein* (2015), and the Showtime series *Penny Dreadful* (2014-2016) extend the franchise and elaborate on its characters.

The manner in which Victor and his descendants have raised the dead have evolved even as science has progressed since the novel was first published. In the novel Victor relied on corpses for his raw materials, but his later film iterations killed for the parts they needed, with others crafting human-animal hybrid and eventually beginning to tinker with genetics. Scores of short stories and films returned the Shelley's well to resurrect through surgery and electricity, with others like H.G. Wells applying variants of Frankenstein's techniques to create human-animal hybrids such as in *The Island of Doctor Moreau* (1896). By the middle of the 20th century film scientists were preserving the heads and brains of accident or murder victims (*Donovan's Brain* in 1953 and *The Brain That Wouldn't Die* in 1962) or experimenting with more extreme alterations such as the flying, bat-winged decapitated head in *Reanimator* (1985). Much more recent low-budget genre mash-ups like *Frankenstein's Army* (2013) use a dieselpunk aesthetic to turn Nazi soldiers into undead cyborgs,to merge flesh and machine into ghastly NeoGothic techno-golems with limited individual individual agency unlike Victor's Creature.

Once science caught up with and even began to surpass science fiction (heart transplants, synthetic hearts, and prosthetic limbs), the question of scientific responsibility turned to issues including artificial insemination, cloning, bioengineering, and genetic modification. Warnings against the social cost of eugenics and genetic testing are central to the dystopia *Gattaca* (1997), while *The Island* (2005) and *Never Let You Go* (2010) ask their audiences to consider clones as humans, individuals in their own right who should not be grown solely to be harvested for replacement parts. Cloning and genetic modification also led to Dr. Ian Malcolm (Jeff Goldblum) giving voice to Shelley's questions in *Jurassic Park* (1993) when he condemns John Hammond, the commercially-driven medical industrialist, saying: "You spent so much time asking if you could, you never stopped to ask if you should!" Returning the dead to life through gene-splicing continues as a theme throughout the Jurassic Park franchise (1993-2015), but returns more directly to the creation of an innocent creature who becomes vengeful in *Splice* (2010). The genetically-constructed hybrid life-form was created as a proof of concept, one that turns on its creators–seducing father, shifting form to rape its mother, and then kill its father, a twist on both the Classical Prometheus and Oedipus. The replicants of *Bladerunner* (1982) are also shadows of the Creature, designed as short-lived laborers, sex workers, and soldiers. As with the Creature, the replicant Roy Batty questions his creator, Eldon Tyrell about why they have been made as they are–limited to brief, five-year lives, before gruesomely executing him. *Bladerunner* asks not just what it means to be human, but also whether replicants could be more human than human.

Asimov used the "Frankenstein Complex" in his stories to not only warn of humanity's tenuous control over our technology, but also to caution against the profit-based replacement of human labor with automation. Continuing to write robot stories into the 1980s, Asimov's greatest contribution to the discourse is the Three Laws of Robotics; he was also proud of coining the term robotics. Although his interest lay in depicting humanoid robots and androids, Asimov's portrayal of man's displacement by machine and inability to control extremely advanced computers also threads its way into supercomputers, both early mainframes and today's networked and cloud-based systems. On a more personal level is the psychological breakdown of HAL 9000 in *2001: A Space Odyssey* (1968), conflicting directives in its programming drive it to murder one human and almost kill the other. While less sophisticated, the trope of government mainframes turning on their master programmers and, later, government "masters" to threaten all humanity has been tied to Cold War fears since the 1970s. *The Terminator* film and television franchise (1984 – 2015) featured Skynet deciding that humans endangered it, so after triggering a nuclear war it created the Terminators to hunt and exterminate the survivors. The Red Queens of the Umbrella Corporation in the *Resident Evil* films (2002 – 2016) turn against man to fulfill corporate agendas, including global apocalypse. But it might be the *The Matrix* trilogy that takes the Frankenstein Complex to its extreme not by annihilating humanity, but rather by subjugating man to the machines, turning almost every human into a fuel source.

It is the humanoid android and the cyborg that blends man and machine that evokes some of the most dire warnings. Some, like Terminators, are closer to robots, limited by their programs with little ability to innovate–that takes something closer to a human intelligence. Others, like Robocop (1986) and Deathlok of *Marvel's Agents of SHIELD* (2015-) must fight to regain their humanity by overcoming their internal programming. Some wear more human shells as a means of evoking sympathy: *D.A.R.Y.L.* (1985), *Bicentennial Man* (1999), and *A.I. Artificial Intelligence* (2001), but empathy quickly becomes entangled with lust and desire as android are sexualized in *Westworld* (1973 film and 2016 HBO series), *The Stepford Wives* (1975, 2004), and the Cylons of the reimagined *Battlestar Galactica* (*BSG*) (2003 – 2010). The Cylons, like Skynet, seek to replace humanity.

The Creature of *Frankenstein* is a tragic figure, one that freely and generously gave food and labor to an unknowing human family. When he tried to join them, he was rejected solely on his appearance–they did not deem him human enough. Science Fiction media and literature has told this tale of creation, abandonment, and rejection again and again, with robots and Cybermen, androids and Cylon lovers, to vastly superior AIs that surpass and replace humanity by transcending them after the Singularity. Beyond SF, online in our social networks and institutional intranets, Asimov's articulation remains trenchant as bots proliferate on social networks, possibly altering elections and democracies, but almost definitely manipulating buying habits, media consumption, and social interaction. Driverless cars are replacing taxis and hired drivers, with the possibility of autonomous trucks replacing another segment of the human workforce. These forms of technology, like mobile computing were meant to enhance and augment human society, but they also alter it in subtle and possibly insidious fashion. These discussions may draw from the more dramatic scenes of popular culture, but will engage the issue of human cognition and the role of technology as it must be considered even as innovation overtakes caution.

1

Design Ethics in the Age of Machines

David Danks

David Danks is the L.L. Thurstone Professor of Philosophy & Psychology, and Head of the Department of Philosophy, at Carnegie Mellon University. He works at the intersection of philosophy, cognitive science, and machine learning, integrating ideas, methods, and frameworks from each to advance our understanding of complex, cross-disciplinary problems.

Most recently, Danks has used interdisciplinary approaches to address the human and social impacts when autonomous capabilities are introduced into technological systems, whether self-driving cars, autonomous weapons, or healthcare robots. His work is both theoretical and practical, including collaborations with industry groups and government agencies. His earlier work on computational cognitive science resulted in his book, Unifying the Mind: Cognitive Representations as Graphical Models, which developed an integrated cognitive model of complex human cognition.

I am particularly interested in the ways in which our technologies encode and help realize our values.

In a lot of cases, we look at things like artificial intelligence (A.I.) and machine learning and think to ourselves that those are neutral, they're objective, they're algorithms, they're math. But they're actually encoding values. They're encoding commitments to certain things being important, which is whatever the algorithm is trying to optimize or succeed at.

The problem is that oftentimes when we're developing these things, we don't talk that way. We don't think about the values that we are or should be putting into our technologies. As a result, you can get a lot of unintended consequences, not because the system did something we didn't expect in some instance — it did exactly what it was programmed to do — but rather because we didn't think about what we were creating, about what its values were.

As technologies have gained autonomous capabilities, they can start to plan and learn, and even make decisions. That change moves humans out of the planning, learning, decision making processes. We're used to the idea that humans bring their values to the table, that humans view the world through a particular lens. We're now starting to have technologies that have goals, technologies that plan and decide, and we don't quite know how to think about values for those technologies. We don't raise them from birth and inculcate values in them as they develop, as we do with children.

As you get things like autonomous vehicles, autonomous weapons, and healthcare robotics, the values that the systems have — what counts as success, their goals — are getting a lot more complex, and a lot more human-like. And yet we aren't necessarily talking about what goals they *should* have, what values they *should* have.

More and more, we're seeing challenges around values in technologies, both what they should be and how we can learn about them. Take self-driving cars. They are of immediate importance to many of us. People talk about this wonderful (in a certain sense) utopia fifty years from now when every vehicle is self-driving and all of these sorts of things.

I worry about not that far-distant future, but what happens over the next five years as people start to trust these technologies more, as people start to count on them more? How many more cases are we going to have like the Tesla autopilot crash where it slammed into a semi-trailer, partly because the human wasn't paying attention?

The history of aviation shows us the challenges of moving between the machine controlling itself — for example, the autopilot — and the human flying it. In the Air France crash a few years ago, that was exactly the problem: a miscommunication between the machine and the human pilot.

We're entering a very tricky time where machines are starting to have capabilities to do things where we've always previously said, "Look, the human will take care of that part," such as deciding if this is a good time to use cruise control.

Or deciding if this is a place that I can actually squeeze my car through, rather than the collision detection saying, "No, stop it." We've always trusted the humans to do that, but now we're starting to offload that onto machines. And we don't quite know what we're offloading. We don't know the limits of our machines' capabilities, and in particular, what happens when we roll it out to the public.

My colleagues in robotics actually have a really good understanding of what their systems can and can't do. And I spend part of my time doing A.I. and machine learning myself. I know how my algorithms are going to work because I built the algorithm. I've spent months living with these algorithms in my head, trying to make sense of them and doing all of the contingency planning and hypotheticals.

The problem comes when I hand that algorithm off to my teenage daughter. She won't have a clue what it's going to do and I think that we're seeing some of that happening right now. More and more algorithms are being used by people who have no understanding of what's going on with them. It's creating problems because they don't know how to interpret the algorithms or their output. They don't know the values of the system, or the values of the technology.

Design Values: Follow the Law or Drive Safely

I want to briefly say something about self-driving cars because a lot of people talk about The Trolley Problem[1] as the place where the values come in. There are engineering challenges to build a car that can successfully drive on the road, but they think ethics comes in only because of these moral dilemmas, these cases where there is no right answer.

I want to push back and say it's a mistake to think that way. Should your self-driving car follow the law or drive safely? That's not a Trolley Problem.

[1] The Trolley Problem is an oft-cited ethical dilemma used to highlight the problem with A.I. The scenario is this: A trolley is barreling down its track. There is a group of people, unable to move, in the middle of the track. You are standing by a lever that would allow the trolley to switch tracks, thus missing the group of people. However, you notice that there is one person, unable to move, on the second track. The dilemma: who do you save? The group or the individual?

That's not, "Do I kill 1 or 15?" It seems like just part of the engineering challenge. But it's actually a value challenge because you can't always do both.

If the traffic around you is exceeding the speed limit, the safest thing to do is drive with the traffic. If it's a twenty-five mile per hour zone, but everyone's doing forty-five, then the safest way to drive requires you to break the law.

Which one are you going to do?

You can't even put the car on the road unless you've made some decision about that.The developers at Uber, Argo, and Google had to make these decisions. We know in some cases what they decided.

Google decided, "We're going to follow the law. Our car will never go more than five miles an hour over the speed limit." You can be in a twenty-five mile per hour zone, everyone's doing sixty, and it will drive thirty miles per hour[2] even though that is actually unsafe. That's a value choice they made.

I sometimes worry that the debates about autonomous cars have a tendency to jump straight to Trolley Problems. You get people who push back and say "Come on, those are so rare." I'll speak for myself. I've been driving now for well over half my life. I've never encountered a situation where I had to make a choice like that. I've found myself in some dangerous situations, but there's always been an escape route.

My point is you're doing values just to get the thing to drive a block. You as an engineer, as a developer, have made a value decision and that's OK.

Sometimes, the engineers and developers get scared by words like 'values' and 'ethics'. They say, "I wasn't trained in that. I don't know how to do it." I say, "First of all, there's a lot of people over on this side of campus who do work on this. Also, you're a human. Humans have values. We want things. We understand goals and success. So you already know how to think about values."

Invisible Decisions A.I.

Where I think this gets even harder is when we start to think about issues of pervasive, ubiquitous technologies, and invisible decision making. The decisions that the Uber cars make are very visible. In fact, it seems as though

[2] Typically, police give drivers a five mile per hour cushion.

Uber has programmed them to be very conservative, in part to avoid accidents because of how visible their actions are.

The big five tech companies — Amazon, Apple, Google, Facebook, Microsoft — are doing an enormous amount of decision making on our behalf that we don't notice.

If I type in certain strings into Google Search, I'm going to get different results than my daughter does, or than you. On the one hand, this is great. If I type in an ambiguous name like Matthew Smith, it's more likely to show me the philosopher, Matthew Smith, than some other random person. If my daughter did it, it would show *Doctor Who*. The exact same string is going to lead to different search results.

That's amazingly helpful. Names are sometimes ambiguous. I really appreciate that I can type in a somewhat common name and I'm probably getting the person that I want somewhere in the first few hits on Google.

Of course, that also means they are making decisions now. They're deciding what to show me. Especially in the time of political polarization that we have, everyone likes to talk about the echo chamber of Facebook. The echo chamber of Google isn't being talked about.

Or consider Amazon, which could be mining all the passive data collection from Alexa, from the Amazon Home, to shape future searches that you do. If you are talking to a friend over dinner — so not the device but you're talking to a friend over dinner — about thinking of going to London, potentially they could show you rain jackets the next time you're on.

That seems wonderful. It does reduce cognitive cost. It is more efficient in your own time and energy, and I know I'm overwhelmed. I think most people are. I really appreciate that. We as humans are not cognitively built for as complex a world as we have managed to create.

Of course, we want to offload things. It's particularly worrisome when we don't realize what we're offloading. We don't even know that these decisions are being made. Even if we wanted to, there'd be no way to do the archaeology to find out why the decisions were being made.

There was a time when Amazon would make recommendations, you could click “Why are you recommending this to me?” My prediction would be that’s going to be become less and less common.

I can’t ask Google why it shows me the things it does in the search. They’ll say, “PageRank,” and I go, “OK, that’s not helpful.” They go, “Yes, exactly. That’s our competitive advantage. It’s precisely that we’ve got this technology that can figure this out.”

I think there’s no question that the Google engineers couldn’t (reverse-engineer) every decision. I think that’s part of the reason that there’s now been a significant push over the last couple of years for things to use — the terms that get used are Explainable A.I., Intelligible A.I., Transparent A.I.[3] — precisely because we can’t always explain what’s going on.

If I don’t know why the car made that decision, if I don’t know why Google recommended these things, then that’s a problem because we expect the algorithms don’t necessarily have exactly the same goals that we do.

Let’s call that bias. We can actually counteract bias in certain ways as the human end user, if we know it’s present. If we don’t know it’s present, if it’s invisible, then we can’t do anything about it. As so there’s been this big push for explainability.

First, Explainable A.I. is a subset of all possible A.I., which means we may well have to make a choice of whether we’re willing to accept worse performance in return for explainability. Are we prepared to say, “Google might be less likely to show me the things I care about, but at least I’ll be able to find out why it showed me what it did”?

That’s a tough choice. More importantly, Google’s never going to do that.

There’s a technological or empirical question: How much is the tradeoff? How much do we actually lose in performance by requiring explainability? But there’s also the question: Is this one of those things that people say they want, but don’t really?

[3] The Defense Advanced Research Projects Agency (DARPA) has a dedicated group working on this problem because "the effectiveness of these systems is limited by the machine’s current inability to explain their decisions and actions to human user."

At that point, we then have to ask as a society, as a political body, “Are we going to be paternalistic?” There are lots of things that people act as though they want that we say, as a society, “I’m sorry, we’re not going to let you have that.”

We can play an active role in regulating these invisible decisions. We don’t have to sit back passively and wait for the technology to run over us. That’s a discussion we have to be willing to have.

What’s in an Explanation?

Another problem with Explainable A.I. is often what counts as an explanation is when somebody with a Ph. D. in computer science can figure it out, if you give them a day and tell them what was going on.

A.I. isn’t a complete black box, but for most people it functions like a black box. Researchers say, “Oh, it’s explainable.” You say, “What does explainable mean?” “Oh, well, you can do a trace in machine code of what happened.” That’s not explainable.

We’ve got a long history in cognitive science and philosophy of trying to figure out what counts as a good explanation. You might think that would be something that would be useful to bring to the table for this debate. “OK, maybe what we should care about is A.I. that’s explainable for everyday people.”

One of the things that we know from psychology and philosophy is that explanations depend on your background knowledge. What counts as an explanation for me might not be one for you (and vice versa) because we know different things.

Now, we’ve got the hard question of figuring out, inasmuch as there is a single public, what does the public know about technology? And what do they believe about technology that’s probably false?

We know that humans love to anthropomorphize things. We anthropomorphize our animals. I’ve got a dog. Most of the time, I think of her as having beliefs and desires, and that works just fine. Every once in awhile, something happens, and I say, “Oh yeah, that’s right. Her brain does not work like ours does.”

We’re entering a similar phase arguably with smart technologies, especially as they acquire certain human-like characteristics, such as the ability to convert words in its memory into acoustic signals.

This isn't speech in the same way that we actually generate speech. It looks like our speech, but Alexa doesn't generate speech like we do. Alexa generates essentially a text string like you would print on a paper or display on a screen, and then reads that to you.

Humans generate thoughts. The expression is a much looser relationship between the thought and the words that are expressed than the thing that's happening with Alexa.

These technologies are acquiring the traits that make it really easy to anthropomorphize them. This actually ties back in an interesting way to the Explainable A.I., because one way to have Explainable A.I. would actually be to say, "You know what? We're going to require our robots and our A.I. systems to act in ways that humans can act, which means interpreting the world as if it's composed of other humans."

You might think, "We need to adjust Alexa so that people can interpret Alexa as if she has beliefs and desires. Alexa wants to help. Alexa wants to give me information. Alexa's trying not to lie to me."

These might actually be ways to generate explanations. The tricky part becomes, "Are they accurate explanations? Where do they break down?" What are those cases like when my dog does something, and I look and say, "Oh yeah, that's right. You actually aren't like me"?

The Commercialization of New Knowledge

These are a set of questions that we're just now starting to ask. And here I have a real concern that the technology is moving so fast, and we don't yet have a solid infrastructure, whether at universities, public policy debates, congress people, industry leaders.

We don't have this in place. We don't have a way right now to have these public debates about, "What kinds of technologies do we want?" In some sense, we're ceding a lot of control to the technology developers. They're saying, "I want this tech, so surely everyone else does. I'm going to do it."

This is an interesting connection with the story of Dr. Frankenstein and Frankenstein's monster. The public — the town folk, in some sense — didn't realize what was going on. They ceded the decision making about their future to Dr. Frankenstein.

One challenge that we have, and it's a challenge we have especially here in universities, we have an ethical obligation to try to help create intelligent discussions about the future technologies that we want as opposed to sitting back. We shouldn't simply react to technological developments, but proactively try to shape them to say, "Here's the technology that we think we want."

They're hard discussions to have. You get pushback. I get pushback when I say things like that to A.I. researchers. They say, "I'm just trying to solve cool problems, and it's not my job what gets done with it." To which I say, "That sounds a lot like the people who came up with nuclear bombs and said, 'But it's not my job where it lands. That's up to the politicians. All I'm doing is creating these things.'" I say, "That's an abdication of responsibility all around."

Your job as a scientist or as an academic is to push the boundaries of human knowledge and to create new understandings of the world. There are many cases where that still works. At the same time, it's one of these things where there are multiple moving pieces here. This is all taking place against the backdrop also of really deep changes in universities.

The push to commercialize the technology and science that's developed at universities has become really quite extreme in some cases. There are universities where the worth of something is judged almost quite literally by the dollar worth of the patents that result from it.

It's actually shifted at many universities away from the pure focus on knowledge creation and towards a focus on being a lot more like industry. At the same time, especially in the computational technology fields — computer science, A.I., robotics — the academic/industry barriers have become very permeable.It's become very common now that somebody will get a PhD, go to work for Google for four years, then go be an assistant professor somewhere for three years, then go to work for DeepMind for two years, and then come back to their university.

There's a lot more movement now, especially in these fields. it's becoming more common for people to think about a professor position as just another stop along your career trajectory, as opposed to the more traditional model where you'd spend 40 years at an institution, for example.

Personalizing the Technology

We still have this mental model of, “Well, if it’s happening in a lab on campus, that means it’s research. We don’t have to think about the impacts. It’s only once it goes across the street that we have to worry.” But it’s not going across the street anymore. The technology development is happening in that lab. The deployment is happening in that lab.

The speed of technological development and deployment has absolutely accelerated in large part precisely because the technology is not particularly hardware-based anymore in many ways.

Like so many things, the deployment of new technologies is going to be dictated by the companies that put out the technologies. The companies can decide, “We’re not going to put out certain kinds of technology,” but they don’t.

It will also be decided in part, and here this always lags behind, by things like regulation.

How do we regulate autonomous vehicles? How do we regulate privacy aspects of cell phone records? If Google Maps is going to track the location of my cell phone even when I’m not running Google Maps, can that be subpoenaed by a police department?

It will be decided by the general public suddenly discovering, “Oh, that’s what’s happening,” and saying, “I don’t want that, so I’m going to choose not to use it, or to turn it off.” And that happens when people have a vivid story that they can imagine happening to them.

I find, for example, the adoption of home healthcare robotics for elderly citizens is still very slow. There are companies starting to build these robots but they’re not in any widespread use, at least in the United States. Actually, Japan is where they’re seeing them come first.

People go, “Yeah, it’s something we should worry about at some point, but maybe let’s wait and see.” There’s a tendency for people to be willing in the abstract to wait until there’s some horrific accident involving a robot accidentally — or deliberately, in the sense of it didn’t know enough to know it shouldn’t do something — killing an elderly person.

When I talk to people, if I say to them, "OK, I want you to think for just a moment. What would it take for you to trust a robot to care for your parent?" By personalizing it, people suddenly realize, "Oh, well, I mean it's OK to have it for other people, but I don't want them." You say, "OK, why not? What is it you're worried about?"

By personalizing these stories, even when they haven't happened yet, it's possible to get people to think seriously about, "Oh, here's what we need to be thinking about now."

I feel like I should make sure to be really explicit and say that I'm actually an optimist by and large about technology, but I try to be a pragmatic optimist. I recognize the challenges, and I realize that we need to be proactive. That's actually the source of my optimism. There are ways for us to try and shape these debates, to shape how these things evolve, to look at the usage of these things by our kids, and say, "No, we're going to actively try and help them become the kinds of people who can use the technology to advance their interests as opposed to becoming slaves to, or dependent upon, the technology."

That requires work. I'm not going to claim that it magically happens, but I do think that there have been enough events now that people are aware that technology is not necessarily going to get us to a utopian future.

The question becomes, "All right. What do we do? How do we shape these debates? How do we get involved in the discussions so that it isn't Dr. Frankenstein off creating the monster without us having any idea what's going on?"

I don't want to suggest that there's no problems or challenges here. It's a really hard challenge. It happens at many levels.

To what extent do we trust people to know what they really want? That's hard. People struggle with that every day. I want to eat chocolate every day. I don't allow myself to. This is a challenge. Part of it is the broader social question. To go super big picture, how are we making space in people's lives so that they can figure out what they actually want?

How do we teach the next generation of developers, even the current generation of developers, or the current generation of CEOs who are doing the strategic directions for their companies that they shouldn't be scared of asking these questions? They're not somehow stealing power from everybody else by asking serious questions about the possible uses or abuses of their technology.

The Shape of Things to Come

It's stunning to talk to CMU undergraduates, and the things that they are completely unconcerned about when it comes to privacy. I'm older than them obviously, but I'm shocked sometimes at the things they're fine with.

This is one of the things that I actually personally find really fascinating to bring it back to what prompted our talking: *Frankenstein*. Some people read that as a completely dystopian book. I don't. I read it as a mixed story. The world is a messy, complicated place. It's hard to know in advance what is going to happen when you do things. Ideas don't always turn out as you planned. You have to have contingencies. You have to rethink things.

Emotion matters. Why does the monster react as it does? In part because, at the moment of its birth, Dr. Frankenstein is horrified and runs.

In particular, it's not that Dr. Frankenstein thought, "Gee, I want to make this creature neurotic." It was an instinctive reaction. How do we respond to these things?

My own view is the future is not a utopia. It's not a dystopia. I was just on a panel last week where the title of it was "Skynet or Shangri-La." But it won't necessarily be just one or the other.

By and large, the technology is built by humans. It's used by humans. It's regulated by humans. There's a lot of these discussions, everyone wants to shift the responsibility onto somebody else. The technologists say, "We just build the tech. It's on the public to decide what they want to do with it."

The people and the everyday public say, "Well, it wouldn't be put in front of us unless it were a good thing to have. The government would stop this if it were bad, or the technologists wouldn't have done it. We can just use whatever gets put in front of us."

The government says, "We don't quite understand these things, and regulation is hard. We've got lobbying from companies, and no lobbying from the other side.It's not our job. We don't know enough yet." Everybody wants to shift the responsibility elsewhere.

What do we do in response to that? Well, I suppose that, we could just be resigned and say, “Well, Skynet: here it comes,” or “Shangri-La: here it comes,” but it’s just a roll of the dice. I’m just optimistic enough, though, to say, “Let’s see if we can try and do something.”

Yesterday, I was at the United Nations talking on a panel in front of a bunch of diplomats about algorithmic bias and autonomous weapons, and trying to help them. It wasn’t to push any particular position. It was trying to help them understand how to think about these problems because, if you’re the ambassador to the U.N., you probably don’t have the background to understand what A.I. really is.

On every dimension: pick your favorite measure — proxy for societal change, attitudinal change, change in norms, change in uses of technologies — and industrialized Western societies are changing faster than they ever have at any point. It’s not just an illusion. It really is happening.

Sometimes, it’s overstated but there is, I think, an important sense in which this is happening. But at the same time, they’re almost no new ideas. What there is, is intelligent recombination, and repurposing, and generalization, and transfer of existing ideas.

When I think about my own research, part of it was taking ideas from machine learning and saying, “Maybe we can make sense of some human behavioral data and psychology using this mathematical framework.”

I didn’t invent the mathematical framework. I didn’t invent the psychological experiments. What I did was go, “I think this thing over here might be usable over there.” You do the hard work to show that yes, it is actually usable, but I think that’s just true over and over.

The history of ideas and most of what we now would call innovations, or discoveries, or brilliant ideas, is the history of recombination. It’s really easy to trace back the history if you just look at them. Charles Darwin did not invent evolution by natural selection. The pieces were all out there. He figured out how to put them together.

The genius is being able to figure out, “OK, here’s these two things that we never really thought about putting together before: a camera and access to the Internet.” Hey, what could you do if those things work together? How can you think about, “It’s more than just live streaming cameras”?

I’m optimistic. I have to be, because otherwise I would just throw up my hands and say, “All right. I’ll go back to just doing the research that advances the boundaries of human knowledge and forget about trying to engage with the public about anything.”

2

A.I., Machine Learning, and Human Augmentation

Jeffrey P. Bigham

__Jeffrey P. Bigham__ is an Associate Professor in the Human-Computer Interaction and Language Technologies Institutes in the School of Computer Science at Carnegie Mellon University. His research combines crowds and artificial intelligence to make novel deployable interactive systems, and ultimately solve hard problems in computer science.

Many of these systems are designed with a deep understanding of the needs of people with disabilities to be useful in their everyday lives. Dr. Bigham received his B.S.E degree in Computer Science from Princeton University in 2003, and received his Ph.D. in Computer Science and Engineering from the University of Washington in 2009. He has received the Alfred P. Sloan Foundation Fellowship, the MIT Technology Review Top 35 Innovators Under 35 Award, and the National Science Foundation CAREER Award.

At some high level, people are really interested in the progress of science and how that impacts people. Right now people are interested about how artificial intelligence (A.I.) is developing and how A.I. is going to impact people.

There's a lot of different perspectives on this. People are really worried A.I. is going to take over, or A.I. is going to put us all out of work, or that A.I. has all of these terrible qualities. On the other side, people really hope A.I. has all this great benefit. That it's going to make all the mundane tasks so much easier so we don't have to worry about that, and we can be free to do more interesting, creative work.

Both of these have some elements of truths. There is a middle ground where there are going to be interesting advantages. There probably won't be A.I. — at least in the near to medium-term — taking over for people. Instead, we'll have more A.I. working with, augmenting, amplifying people for good or bad.

In terms of right now: machine learning — especially understanding patterns and then being able to predict from new incidences of data what's going to happen — is being applied to everything.

We are generating huge amounts of data because we are now interacting with devices that are not only connected with each other, but also can record data and store data. Then we can take that data, aggregate it into these huge training sets, and use that make the machine learning work better. We're applying that concept everywhere. Every website basically that you go to is — in some way — informed by people who have gone there before and the actions people have taken.

It's being used in predicting what you're going to buy, influencing what you're going to buy, deciding what ads are showing. We hear this talk about the election where essentially A.I. was predicting who were the swing voters? Who are the people that we can most influence and how can we micro target those people? Machine learning and A.I. are being used for hiring, insurance decisions. It's being used in all these different places.

There's a lot of benefit that can be done. If you have the machine sort through the data, maybe you can understand patterns that people just manually looking over the data wouldn't. That would allow you to predict things better. That's where the power in A.I. and machine learning comes from.

The danger is that once you have this big complicated system that learns these things automatically and building these models that you can't necessarily understand, it's really difficult to know how the system has made a decision, or recommended a decision, or helped you make a decision.

What problems might arise from that?

Biases maybe encoded in the data. People talk now a lot about how there's A.I. that's amplifying the existing human bias. If you have some bias in your data set

— if there's some human bias that went into creating the initial data — then it's going to be reflected in the machine learning models out there.

This is a very simplified example of having bias. It's probably more complicated in practice.

Let's say that you look at a huge number of hiring decisions that have been made. Then you look at all of the characteristics of the people who were candidates. What you want to do is make it so it's easier for the hiring manager to make these decisions.

You decide to just look at past data — who was hired and who wasn't — as a training data. The model will be able to predict what that hiring manager would decide if they were to look through all the *new* applicants. That's helpful if you get way more people are applying for positions than you can possibly look at. It would improve efficiency. It would mean that we can look at all the candidates, and not just the few that happen to have a personal connection to somebody in the company.

The challenge is this: What if the hiring manager who had been making those decisions was in some way making that decision not just on the characteristics that you would want, but maybe something else? Maybe they're prioritizing males over females, maybe they're prioritizing somebody of a particular race or religion or back, or whatever it is.

But bias gets more complicated.

It may not be that race or gender is specifically encoded in the data that the machine learning algorithm is working on. Still, machine learning is really great at detecting patterns. Maybe you don't even have the gender of the person, you don't even have the race of the person, but you can infer from something about their prior work history, where they got their undergraduate degree, who knows what it is. You can infer something that ends up being correlated to race or gender, or these other things that I think we as a society — at least hopefully — believe are not what we should be hiring people based on.

That's just one example that ends up being the algorithm learning bias from data or is making decisions based on something that maybe isn't what we wanted to be making decisions on, and then that bias gets encoded in the model in a way we don't necessarily see or understand. There may be nothing in that model that says *only hire men*, and it probably wouldn’t even only recommend hiring men.

But it might have a slight tendency to favor male candidates or other candidates, or whatever the bias is that it might have learned.

Unpacking the Black Box

It’s really hard to unpack those algorithms because it’s probably not learning explicitly on dimensions like gender, like race, like the things that we might want it to not be factoring in. Instead, it’s correlated attributes that are really hard to learn.

Making it even more difficult to understand is that the company’s the develop these tools see their models as their secret sauce. They aren’t necessarily going to publish these models in a way that independent third-parties can poke at them to say, “Let’s see what happens if we just send in one-hundred African-American male candidates, and see how does the algorithm makes its decision.”

One reason companies are not going to want you to do that is that you’d probably find something. But the other reason is that with all the different industries using machine learning — from the insurance or finances or whatever it is — this is their competitive edge.

That’s one problem.

The next level of thinking that people get to is where they start to think, “Well, OK. We might not be able to inspect the model and understand this, but we also can’t really inspect the model that a human is using.”

What we do with humans is we judge people based on their outputs. The trick there is if we don’t have access to the model (the decision-making process) and we don’t even have access to the outputs (all the candidates who may have applied) it’s really difficult to reverse engineer what that model might be doing.

You’re left with looking at examples instead. You have a candidate that maybe was or was not hired. You might be wondering, “Well, I wonder if one of these things we don’t want to be considered was actually considered.”

With one or two examples it’s really hard to do that.

But there is this incorrect assumption that the biases that machines learn will map on cleanly to what we know about human bias. Human bias comes in different forms, but we all have the sense of, “Wow, there’s some really terrible kinds of bias that people actually have, and we might wanna look out for that.” Things like racism and sexism.

A machine isn’t a human. It’s not going to necessarily incorporate bias even from biased training data in the same way that a human would. Machine learning isn’t necessarily going to adopt — for lack of a better word — a clearly racist bias. It’s likely to have some kind of much more nuanced bias that is far more difficult to predict.

It may say come up with very specific instances of people that it doesn’t want to hire that may not even be related to human bias. But for whatever reason, its experience with candidates from a particular zip code were never hired. Now the machine receives a new candidate from a particular zip code and because of weird training data that the system has, those candidates will never get through.

The big question that we should be grappling with is this: Should we be pushing this far and this fast?

We’re not grappling seriously enough because so much of what’s going on is hidden. People don’t really understand how much A.I. and machine learning are being used. It’s a huge complicated mess if you have A.I., you have people, and you have these processes that combine both into decision-making models that are very difficult to understand and very difficult to reverse engineer even for the people who are involved in the system.

So this thing that we have rushed into then presents us with this really hard problem. We’ve rushed into simply amplifying humanity with A.I., which essentially what’s happening. But we’ve amplified *everything.* It’s not like racism didn’t exist. It’s not like sexism didn’t exist. All these things always existed.

Before we were mostly limited to one-on-one or small-group communication. But now you’re Uncle Joe can share his conspiracy theory about the Obama presidency, and the algorithms are learning that “Hey, people really like this stuff.” The machines then spread it further and further, well beyond Uncle Joe’s normal audience.

But it’s a mistake to present this bias as easy to understand. Depending on your perspective, bias can mean a whole bunch of different things. What does it even mean not to have bias? How do we control it?

Bias, Connectivity, and Amplification

There’s a couple of things that you could consider on the table of options when making an algorithm. One thing that seems really important is making it so that more than just the people who are writing the algorithms and controlling the data can start to probe at what that bias is.

One way to do that — which would be a huge undertaking — is opening up the models in the data for anything that we see as at all important. One of the reasons why people send a lot of their frustration at Facebook is because really only Facebook knows what’s going on.

People can uncover bits and pieces of the model, but it’s really hard because you need a lot of data before you can start to build up a pattern. If you had the model, if you had the data that it was trained on, then you could do that much more easily. You can imagine — and there would be a lot of resistance to it — policy that basically said if you are a data-oriented company, like pretty much every company is now, then you would have to open up your data and open up the models.

It sounds crazy to say that if you are a company that makes innovative new products, we need you to register those products and those inventions with the government agency.

But we do that now, right? That’s patents, right?

The reason it’s difficult from a policy stand point is because there’s a huge amount of money that is being made — and can be made — from the closed models. Maybe you can do it in a way that respects that. Maybe it’s more about third party audits. But the really difficult challenge in opening up the data is that if you actually expose all of Facebook’s data, even trying to respect people’s privacy, there will be all kinds of things leaked out of that.

Our voting machines are completely vulnerable to attack.The only saving grace has been that each precinct — each little local district — runs its own voting machines. They are not usually networked so you have to go on site to hack them. That’s not true with Facebook. Facebook presents a public API for you to hack Facebook — that's the ad network. We know how to get access to Facebook's algorithms and people are presumably always trying to do that.

This is just one example, but if every company has some sort of interface to their data and to their model, it's not inconceivable that there will be different ways that people try to manipulate it. And we might not be able to detect the people who want to go in and manipulate those algorithms for various purposes.

One thing that is important to point out is that this is also about connectivity. The connectivity that we have now opens up the possibility for A.I. to amplify things. Today, a niche group — and people have talked about this a lot — can now find each other online and organize around something that may be positive or negative.

That makes a very different kind of society than existed twenty years ago.

We don't know what the implications of that outcome may be. It certainly seems like mainstream society would have squashed many of these negative ideas we see online, but now they are finding a foothold because you can connect with people like you. You can have this weird, racist view and then you can find plenty of people online that share that view and then we can go organize. Now, we're in Charlottesville.

Of course, this amplification can have really good benefits, too. A gay teen in a small rural town can now find people to connect with. They can find people like themselves, and know they can have a great life. There are great things with this connectivity.

All that stuff happens and it's not clear what we should do. But it's certainly something that I think we should be talking about.

To the extent that A.I. amplifies these negative ideas in ways that are difficult to understand, we should be working to improve our understanding of how the machines work.

I don't think you can do that just by observing after-the-fact the outputs. You need to have some kind of way to look at the models and look at the data. Maybe that can be done within the context of the corporations who are controlling a lot of this, but I'm actually pretty skeptical that you can really do that without third party review.

I think Facebook is finally feeling some pressure, for instance. Twitter is finally feeling some pressure, and other social media companies feeling some pressure.

Without that pressure, the way that you would naturally model your success on a social media platform is just: Are people viewing the platform? Are they visiting it? Are they staying around? Are they adopting it? Are they looking at my ads?

It's not clear that those metrics of success map onto something that we would think is positive particularly when you move these things algorithms into areas that impact real lives. There's also this human level question: Regardless of the performance of these algorithms, do we want to use past data to influence outcomes and outcomes for individuals?

We seem to be more or less comfortable with this sometimes. For instance, if you're buying life insurance. People have talked about how this has been used for prison sentencing and parole because if you can predict recidivism, you can predict how likely this person is to basically come back. That might influence a final decision.

The problem is, first of all, that you have all this bias. You shouldn't diminish that. The second problem is that even if you can predict recidivism because twenty people who were just like me committed another crime after being released from prison, what do you do about it?

I'm not those twenty people. I might be similar to them on all these kind of easy to record metrics, but I'm not them.It seems somewhat at some level unfair to penalize me because based on the thirty features you can record, I look like those other people.

If you're trying to minimize the overall recidivism rate of the district you're in, then it might seem very obvious.Once you start making statistical arguments about individuals, it gets really tricky because there might be some feature about me that's not captured in the model that makes it more or less likely that I'm likely to come back.

This gets super tricky.

We need to start making explicit decisions about using machine learning to amplify our decisions because if we don't, these models will get deployed. If you look at the metrics by which people are judged, those are the same metrics that people will use to build models. If their metric is "We should not be paroling people who are going to commit a crime," then that's what the model will try to maximize. It's not going to consider the individuals. And if people are not given the leeway to make different kinds of decisions, lots of problems can happen.

We need to think about as a society, are we comfortable with making statistical judgments about individual people? I think everyone needs to understand how these systems works before this becomes a big problem.

Adaptive Technologies

I think there's a huge potential for A.I. and well-designed systems to make people's lives better.

Many of the systems that we built have done things like interpret visual information for people who are blind. We built an app[1] where you would take a picture, then ask a question about it, and you get an answer back within about thirty seconds.

Initially, we had people answering those questions. We had people recruited out on the Web. They were paid to answer these questions, but we were soon able to do more with computer vision. A computer that is able to do that would see the image, add a question, and then would able to send you an answer back, all automatically. The more you can do computer, the faster it can work and the cheaper the service.

Those computers were trained in very similar ways to the ways. You have a lot of data and lots of images and questions and answers. You send them to the machine learning algorithm. They are able to build up a model that's able to predict what the answer will be to something they've never seen before, which is pretty fascinating, right?

To some degree, I think this sort of thing should be the fundamental type of problem for A.I. What's more fundamental to artificial intelligence than trying to reproduce the human perceptual system? That's super cool.

Much of our work is still powered by humans with some A.I. systems but then there's a great deal being done completely automatically. If you go onto

[1] VizWiz allowed sighted people to download the app, and answer questions posted by blind users. The users would speak their questions and answers using their iPhone, allowing for a near real-time exchange of information.

Facebook and you right-click on an image on Facebook — like some image your friends have shared — there will be a little description there that's produced completely automatically by the computer vision system that Facebook has.

They're not always right, but it does work pretty well to tell you how many people are in a photo, if they are smiling, and other information like that. That is intended to provide a more accessible experience for blind web users.

The goal with a lot of A.I. research for thirty or forty years has been to give people systems that augment or amplify their abilities. One common scenario — and Google does this pretty well — is that I can look at some non-English text but I see that text in English through my augmented glasses.

I think it's really clear that what you want to do for people is to give them augmented technologies. You want to give people clues to help them make decisions.

But one of the fascinating things that I've learned working in the space is that it's incredibly difficult to compete with the cane that a blind person would use. The cane is amazing. Somehow the proper reception in your hand, the tactile feeling of hitting the floor, and the sounds it makes provides a really good experience.

It's really hard to replicate that with any sort of technology.

A lot of what I've tried to do in the adaptive technologies space — and what I think is really interesting — is to give people some of the non-obvious clues that some of the people have for interacting either with information or in the world.

One example: You're going into a building you've never been into before. There are always different clues that you use to figure out where you're going. Some of them are very explicit. You might look at room numbers or signs. But often it's pretty subtle. You see a brighter hallway on one side that might orient you to say, "That probably is the way I want to go." You see a door that looks a little different than the other doors and say, "I bet that's the stairwell I need to go."

There are all these little clues that you can access that you may or may not work every time, but you build up and accumulate these implicit knowledge of how you might want to interact with the space.

Getting that kind of information through a piece of technology is hard.

But people with disabilities are often the early adopters of A.I. technology. The utility trade off for them makes sense earlier. I know people who are unable to type on a physical keyboard who have been using speech recognition for twenty years even though speech recognition then was terrible. It was awful. Now, it's finally getting better and people are starting to use it, but people with disabilities have been leading the way as users of this technology.

This is the same thing that is happening with computer vision.

Now with Google Translate, we can go walking around in foreign countries and see the signs converted to our natural language or our native language. I'm a little skeptical about how that will work given that they have to get both speech recognition in the world and translation working all in sequence.But I think the cool thing is that we'll get to see how well it works.

There is a space for this adaptive, A.I. technology to be out in the world there even while it's imperfect. If I think about going to a country where I don't speak the language and if I could just get bits and pieces translated and presented in a good way, I might actually be able to get through an interaction.

We do that today without technology. We may barely understand what the other person is talking about yet somehow it sort of works.

I actually had a deaf colleague who is a sign language speaker. She was in Italy, and found herself at this hotel and there was an Italian hotel owner. There was also somebody from the United States who only spoke English.

She described this weird surreal thing where she became the translator, despite not being able to hear either one of them, because she had a much better familiarity with gestures and the various ways we communicate nonverbally.

That was a really interesting story for me to think about. Really interesting.

That's an example where the speech recognition equivalent didn't work at all. The system didn't work at all, but the other clues and the ability to augment their communication with each other, with gestures and with kind of understanding people did work.

That's a human example of the kind of augmentation that you might see A.I. get to before it's perfect. Because it's going to be a long time before it's perfect.

What those technologies can do — and I think this is the important part about A.I. in general — is that the A.I. will be a tool and the person dealing with the context will be the user. I think that it's going to be a super-long time before A.I. truly understands any of the rich, complicated, complex context that we as humans deal with in real time. It will be a long time until that works, but it doesn't mean it's not useful. It can still be useful because the person using it is the one who understands the context and interprets it.

One of the things people often use the computer vision app we built to do is for prescription bottles. Initially, people were worried about that use. I was a little bit worried. I'm still a little worried about it.

Here's the thing, though. Unless, somehow the A.I. or the human assistant was being malicious, most of the ways that it would get it wrong were obviously wrong.

If you're supposed to take one pill and it tells you ten, it's pretty obvious to you that's not right. If you read the medicine bottle, it's very difficult to confuse Tylenol with some other kind of medicine. The output from the app is more likely to give you some random word or some nonsense. Because of that, the user almost always can make sense of that. And more importantly, they had enough context to know whether it was okay to try to make sense of it.

They had context from their prior experience with the medicine. It's not like this medicine bottle appeared out of nowhere. The user would have some notion that ten pills probably doesn't make sense. There's a lot of context people build into their own lives that the technology doesn't need to solve.

Ten years ago, we built a system that was supposed to help blind shoppers find the product they wanted at a grocery store. We built a shelf. We bought all twelve boxes of cereal. We were all set to try it out. We had to set up this comparison where you took a picture of the whole shelf, and it would use almost a Geiger counter-like model to direct you to the box you wanted. It would just tell you which way to go. It would say, *left, right, go forward.* And we used a barcode reader that would also read out if you found the barcode.

The systems was kind of hard to use. You had to find the right side of the box and scan it so it could tell you what the product was.

We thought let's compare this with how blind folks shopped without the system. We we thought we'd win. Surely it would take them forever for them to find the right box. They'd have to go through every single box. We thought that would take forever. Our system was flaky, but it would eventually get you to the right spot.

Every single participant who came in didn't even touch the barcode on the box. They just went through every box, shook it, figured out which one they thought was likely the candidate, and selected. They were like, "This is definitely the Mini-Wheats. It's the heaviest, and it shakes like Mini-Wheats."

They were right almost every time. Then they would just verify that they were right. It was so fast. It was really annoying because they beat our cool system.

What we learned was that they had context our technology lacked. This is what almost always technology — especially in the assistive technology — in general often ignores. We have to understand what were people doing before we create our cool, new technology.

For a system trying to be intelligent, it's really difficult to compete with your own context or your own intelligence.

Still, I think much of the innovation coming out of A.I. and machine learning is going to be in the adaptive technologies space. There are just a lot of opportunities. If you can design the technology well enough to both have really interesting A.I. capabilities and meet the user needs, there's a huge potential.

There's also a big opportunity for technology that doesn't seek to completely replace the user's own perception but instead works with it. The computer vision is now getting good enough that I think it can start to augment the understanding of the environment for a blind person.

I think that speech recognition and signal processing of sound is getting good enough that it can start to augment what a deaf or hard-of-hearing person knows about their environment. You can probably make other kinds of analogous predictions for other kinds of disabilities.

We're not going to be at the point where you would completely rely on a computer vision system or speech recognition system. There are really difficult problems to solve. The really difficult problem is how do you actually build the technology so a person can benefit from it.

This is what gets constantly overlooked. You’ve got people who are really great at the machine learning that built some model that’s able to do something with a dataset. Transferring performance on a dataset into something that people really want to use is really hard.

That’s where we’ll see a lot of innovation with deep learning. The A.I. has gotten to the point where it’s good enough. We just still figuring out how to transfer it to something that’s actually useful.

3

A.I., Design, and Absurdity

Molly Wright Steenson

***Molly Wright Steenson** is a designer, researcher, and author whose work focuses on the intersection of design, architecture, and artificial intelligence. She is an associate professor at Carnegie Mellon University in the School of Design and author of the forthcoming book Architectural Intelligence: How Designers and Architects Created the Digital Landscape (MIT Press, Fall 2017), which tells the radical history of A.I.'s impact on design and architecture and how it poured the foundation for contemporary digital design.*

A web pioneer since 1994, she's worked at groundbreaking design studios, consultancies, and Fortune 500 companies. She's previously been a journalism professor at the University of Wisconsin-Madison, an adjunct at Art Center in the Media Design Practices program, and a resident professor at the Interaction Design Institute Ivrea in Ivrea, Italy. Molly holds a PhD in architecture from Princeton University and a master's in architectural history from Yale.

I am very interested in what we're saying when we say artificial intelligence (A.I.). I'm interested in the long-term stakes of that term, curious about its history, and a little suspicious of how it gets used today.

In particular I come at the definition of artificial intelligence through design and architecture.

The term was coined in 1955 by John McCarthy with the Dartmouth Summer Research Project on Artificial Intelligence, a summer research conference that he convened to pull together a bunch of people — including Marvin Minsky[1], and Frank Rosenblad, and Claude Shannon — working on a set of things around what would become known as A.I.

He laid out a platform of about ten items for Artificial Intelligence research[2]. These platforms are largely in place today — looking at machine learning, looking at neural networks. looking at games, looking at natural language processing. These various research platforms were put out, and this group of researchers got together for the summer to sort them out.

There were a number of research programs that were pursued that began at the Dartmouth conference. The MIT A.I. Lab, the Stanford Artificial Intelligence Lab that John McCarthy left to lead after founding the A.I. Lab with Minsky. There was Project MAC, which gave rise to time-sharing computers. There was Multics which has a long lineage of the beginning of important programming platforms and connections, eventually to UNIX, and to eventually the ARPANET and Internet as we know it. Different labs worked on different parts of the problem.

It was 1961 that Marvin Minsky wrote “Steps Toward Artificial Intelligence,” outlining what he thought some of the key problems were going to be. Many of the ideas that we talk about in regards to human-computer symbiosis — an idea was first published by J.C.R. Licklider[3] in 1960 in “Man-Computer Symbiosis” — came out of this work.

The parts that interest me are not only the history but also how collaborations with architects like Nicholas Negroponte or architectural applications actually gave rise to A.I. That’s a major part of A.I.’s development. There are interesting crossovers there that happened.

Nicholas Negroponte founds the Architectural Machine Group at MIT in 1967, which is the predecessor to the Media Lab[4]. In fact, four of the founding labs

[1] Marvin Minsky was the co-founder of MIT's A.I. Lab. He was one of the leading figures in A.I. research

[2] There's much written about the summer conference. Grace Solomonoff wrote a paper, "Ray Solomonoff and the Dartmouth Summer Research Project in Artificial Intelligence, 1956", that chronicles the event from its inception through the development of the items that would shape the A.I. field.

[3] J.C.R. Licklider oversaw the funding that would lead to the development of the ARPANet, the precursor to the modern Internet.

at the Media Lab come from what was the Architecture Machine Group at that point.

J.C.R. Licklider was one of Negroponte's mentors and funders. Minsky was one of Negroponte's collaborators, and the Architecture Machine Group used the experiments and projects that the A.I. Lab was doing. They were developing interfaces for Artificial Intelligence between 1967 and 1984.

Blocks Worlds

If you look at the history of A.I., there are a couple of paradigms that shaped how that technology was being developed. I mentioned them because I think they're important for us to keep in mind today as we do A.I. research.

The first paradigm is microworlds or blocks worlds. Microworlds are limited domains in which A.I. researchers focus on a specific problem like using computer vision in a pile of blocks. These are called blocks worlds because they literally were stacks of blocks and robotic arm manipulation. It's a way of zooming into the very small particulars of the problem.

Minsky wrote[5] about the fact that these problems might not be true if extrapolated to the size of real world. You have a way of working on a model and working on these questions of computer vision that allows you to work on a specific problem. As they put it: it's a fairyland. If you were to deploy it at the scale of a city, the problem is false and the solution doesn't work.

There's a project that the Architecture Machine Group did in 1970 called SEEK, which is a block world. It's a set of mirrored blocks and a robotic arm that tries to stack and organize the blocks by following a set of six programs — straighten, stack, etcetera. The project used computer visioning to figure out where the blocks are and stack them.

[4] The M.I.T. Media Lab is a practical, multidisciplinary graduate lab where students work across disciplines — from computer science to design to music to architecture — to build technologies that aid and amplify how we live.

[5] "Each model -- or 'micro-world' as we shall call it -- is very schematic; it talks about a fairyland in which things are so simplified that almost every statement about them would be literally false if asserted about in the real world. [...] Nevertheless, we feel that they [the micro-worlds] are so important that we are assigning a large portion of our effort toward developing a collection of these micro-worlds and finding how to use the suggestive and predictive powers of the models without being overcome by their incompatibility with literal truth." (Internal MIT memo Minsky & Papert, 1970; quoted in Dreyfus, 1981)

This application of the robot arm was in development in MIT A.I. lab and the stacking was a developed with the Architecture Machine Group.

For the software show at the Jewish Museum in 1970, they introduced the gerbils as an art project in this big five-foot by eight-foot pen with 400 cubes. They wanted to show that the computer model and the gerbil model were supposed to come into conflict. And the gerbils did what gerbils do, which is make messes and nudge around the blocks.

That was all well and good except for one thing: Seek tended to kill the gerbils.

This says something interesting about micro-worlds or other bounded environments and problems with A.I. modeling in general. If you want to focus in on something very small and specific, it isn't that straightforward. It's not just a matter that the gerbils are mismatched with the computers. The programmers and designers mismatched what happens when they scaled the model to a much bigger size.

The thing I'd find myself thinking about today is that when we come up with plans or ideas of what happens in the lab and then scale them up to say the size of a smart city or Internet of Things, we're looking at scale and networks, we start seeing some strange results.

Microworlds actually deliver micro-knowledge. They can't scale up. The big promises that A.I. was making just simply weren't coming to fruition.

Command and Control

The Mansfield Amendment also happened in 1970. The history of military funding is really important here and actually inextricable from the history of A.I. research. The Mansfield Amendment was introduced so that defense funding could no longer be used for basic research. It needed to be used for applied research only.

At the time, the National Science Foundation had a very different funding culture. The amendment changed that from "We wanted to tinker in it and see how it works and figure out what the use might be later" to "This has an applied military tactical application."

That is a big shift for basic research in A.I.

These combined effects meant micro-worlds wouldn't be funded anymore. So A.I. researchers turned to command and control because that was the next paradigm in military funding.

Command and control is tactical about battlefield control. You're not talking about gestures and computers.

One thing to know is that the Office of Naval Research was instrumental in setting the agenda for A.I. funding. In Minsky's book, *The Society of Mind*, he thanks a man named Marvin Denicoff, who was an absolutely vital force in setting the agenda for all A.I. researchers at the Office of Naval Research.

Until this time, A.I. researchers tended to work in a closed world. They would pass ideas between each other, and they worked together for a long time. This new paradigm meant longstanding relationships were not valued. Negroponte had explicitly said how much he hated that. I think any number of people who were working with DARPA[6] would have preferred to stay in that model.

But the funding agencies instead started turning their projects toward tactical battlefield work.

You begin to see projects like the Aspen Movie Map in 1979. The Architecture Machine Group researchers went out to Aspen, strapped some movie cameras to a jeep, drove through the streets, put the images they found on a video disk, and made a simulation of driving in which allowed you to zoom through the streets of Aspen.

You could experience this in the Media Room, which is an immersive media environment that the Architecture Machine Group and then the Media Lab had into the early eighties. You can zoom down the streets of Aspen and you have a satellite map and just a basic line map on touchscreens on either side of you as you're sitting in an Eames Lounge Chair that has been retrofitted with pads which were touch-sensitive joysticks.

It was a big screen that showed the streets of Aspen displayed in front of you. It's proto-Google Street View, but the tactical purpose is military surveillance.

[6] The Advanced Research Projects Agency (ARPA) — which would add Defense (DARPA) to its name in 1972 — has long funded technology research and projects that could be used for national security.

This isn't strictly A.I. The command and control research was supporting other technologies that we know of today. There's a long history of virtual reality (V.R.) that comes out of the Architecture Machine Group, but really the person you want to look at for some of the important beginnings of A.R. and V.R. is Ivan Sutherland[7] and his head-mounted display from 1967. Even Ted Nelson [8]was looking at ideas around virtual reality and augmented reality.

There are other projects that may not be artificial intelligence per se but it's funding for command and control made them possible. In this sense, the Architecture Machine Group collaborating with the A.I. Lab and with other bodies at MIT created objects that fell under the command and control umbrella and that were experimenting with different kinds of interfaces.

A.I., V.R., and Architecture Machines

For Negroponte, A.I., interfaces, and V.R. are inextricable. In 1970, he wrote a book called *The Architecture Machin*e and in 1976 he published another one called *Soft Architecture Machines* where he puts forth a theory of architecture machines, in which he describes a future where we won't use a computer.

Instead, we will live inside of that computerized environment.

It's his view of a computerized environment that I think we all find ourselves in in a world where we've been talking about ubiquitous computing. Ubiquitous computing is a term popularized by Mark Weiser at Xerox PARC. It's the idea that computers and sensors sink into the world around us, and the world around us becomes interactive.

We now actually exist in this world.

This idea is a potent idea. Eric Schmidt[9] told a group at Davos that world is going to interact with you and that computers will be everywhere. But the gist of this is actually that the computers will disappear.

[7] A graduate of Carnegie Mellon, Sutherland is often referred to as the "father of computer graphics." His head-mounted system, dubbed The Sword of Damocles, was the first virtual and augmented reality system.

[8] A sociologist, Nelson created Project Xanadu, a project designed to build a simple user interface for computers. He also coined the terms *hypertext* and *hypermedia.*

[9] Schmidt, a former member of Carnegie Mellon's Board of Trustees, is the Executive Chairman of Alphabet, the parent company of Google. He previously served as Google's CEO from 2001 until 2011.

In architecture, Le Corbusier said “A house is a machine for living in.” There’s a long history throughout twentieth century architecture of thinking of our dwellings as interacting with us. Negroponte began to flip that on its head. Everything is a dwelling when the computer disappears.

The question for him is about interfaces. He’s frustrated by the quality of interfaces, of input and output devices, what our bodies do, and what our interfaces do.

He writes: “Does a machine need to have interfaces like my body’s in order to be intelligent?” He noodles on this for a little while and he says the answer might seem absurd. But he believed the answer is yes. Machines needed to have sensors like our personal sensors.

Machines need to have new ways to input information and the outputs need to be richer too. For him, you can’t separate out notions of artificial intelligence from how you experience that intelligence.

This is this kind of melding of man and machine, entertainment, war, all of it. By the late 1970s, they are pitching media spaces and information spaces. They created something called data space and the spatial data management system.

The Architecture Machine Group (Guy Weinzapfel) created the first digital layered maps. You start seeing pictures for augmented reality. You see things that look like an iPad with an IBM Selectric typewriter in the background and a dictaphone. All of these things are necessary to have a sentient environment.

More Human than Human

The questions for me today are: How do we design for for intelligence? Where do these notions of intelligence come from?

They were developed with architects and designers and so I need to look back at those collaborations to figure out what was baked into it to begin with.

We can take the fact that maybe Google wants to talk about machine intelligence. We could talk about augmented intelligence. We could talk about A.I. We could talk about machine learning being an entire field of work and study and research and practice.

Sometimes they all mean the same thing depending on who’s talking about it and who’s wielding the terms.

I keep coming back to the fact that these words are old. When we talk about augmentation, we're actually talking about a question of something that someone like Steven Coons[10] would have put forward in 1963 as he was working on CAD systems. You could take a look at the idea that Negroponte or Gordon Pask, the cybernetician, puts forward that the computer and the human together do something more and better that what either might do alone.

Gordon Pask is saying that each entity can develop. It becomes something else, something other. It suggest that it's not just a matter of equal partners. It suggests things will happen that we never would have come up with on our own. Then you have surprise, you have serendipity, an uncanniness.

I know that we say the nature of this black box is one that we don't understand what's happening. For instance, there's the recent example of Facebook bots developing a language that humans couldn't understand. It's like, "Ah! Shut it down!" Maybe shutting it down was a good idea or maybe it was not a good idea. But wouldn't it be interesting to focus on it in a way that wasn't fear-based?

I think also there's a question computationally of what's going on. Can it be understood?

I suppose you can always find explanations for anything if you look hard and long enough. We've been working on black holes, stuff that Einstein wrote about general and special relativity for a century. We're beginning to bare some of it out. Sometimes it takes a while.

That stuff goes and goes and goes and it seems like old conversations. If I bring this back to questions of design and architecture, that's really productive territory to look at.

Architecture and Design

Gordon Pask worked with an architect named Cedric Price to design what could be called one of the first intelligent buildings called Generator. It had a set of cubes and walkways and barriers, which were all recombinable. He designed this project for an arts retreat center in Georgia, which was never built I should mention.

[10] Coons was a professor at M.I.T. who worked on design interfaces that could be used by engineers. The Steven Anson Coons Award for Outstanding Creative Contributions to Computer Graphics is presented every other year at the SIGGRAPH conference.

But Price wanted people to be able to move these blocks around, and recombine them at will to suit their needs. But he realized people tend not to do that. So he got in touch with John and Julia Frazer, who were computer graphics researchers and architects.

They proposed creating a set of four programs and microcontrollers. Then they had an inventory program that would say where each piece was as well as a set of rules about how those blocks could be combined and how they could not. They had a physical plotting program where you could move around the blocks. It would plot the design on a computer screen and then you could print it out.

The fourth program was the boredom program.

In the event that people hadn't moved the blocks in some set time period, the Generator would come up with its own plans. It redesigns itself because it gets bored.

That's uncanniness. That's the idea of not just human-computer symbiosis working together to get a job done but rather, human-computer partners coming up with sometimes surprising results. The computer is programmed to get bored.

But the question comes up: Is the computer intelligent? No. The computer is programmed to be intelligent. But I think Cedric Price was very interested in architecture that didn't stay still and that wasn't determined. He liked indeterminacy.

I think there's an interesting tension between we want computers to do our bidding and we want computers to surprise us. That question of surprise is really important in the space of design and art and architecture, as long as you don't get people killed because your buildings are falling down.

There's another side to this. The uncanniness is actually what pushes us to think differently about what intelligence is. I'm beginning to think that having a sense of absurdity — more than things having a sense of humor — might be how you show intelligence.

I think the notion of intelligence, in general, should be that it is about getting some plan you didn't expect. That that's a really interesting idea. The unexpected should be how we know that something is intelligent.

It’s the “Surprise Symphony“. Haydn had people falling asleep. Then boom. That’s how you get people to sit up because you’ve done something different, unexpected.

I love Janelle Shane’s work. She’s a neural net researcher. She trains neural nets to do very funny things. Paint colors. She fed seven thousand paint names in RGB values to her neural net and trained it to start working on paint colors and names. They have names like Burble Simp and Stoomy Brown and Stoner Blue and Turdly. It’s just so funny. All the colors are these limp blues and greens and beiges and puces. That’s very funny.

And she does Pokemon names, eighties action figures, death metal bands, and guinea pig names that are sometimes a mix of all of those things.

I really like uncanniness and absurdity. For me, those things stand out about what it tells us about the nature of something. The kind of jokes someone cracks tells you what kind of person they are. If someone around you starts telling racist jokes, you’re like, “Oh boy. I have a very different view of you now. I think I’ve got to be going.”

I tell terrible dad jokes. It says something pathetic about me, I know this. I also think that it’s maybe productive to look at the history.

This is something I’m just kind of working on in my mind. Look at the history of the twenties century art, like the emergence of Dada, right during World War I. What does art mean when the world is going to hell? We needed something that’s absurd to turn it on its head. There was a real push to upend traditional art institutions at that point in time.

Or the writing of someone like Eugène Ionesco right in the wake of World War II and totalitarianism. There’s not an elephant in the room. There’s a rhinoceros that’s coming to get us, and everyone’s turning into rhinoceroses, or the bald soprano, where the fireman is the person who comes and knocks on your door. You need these narratives to up end the expectedness because things are different.

We’re in a pretty in a pretty strange timeline now. It would be nice to see more of the humor and absurdism — the serious absurdism — come to play in what’s unexpected about Artificial Intelligence and design. Something that doesn’t lead us to shut down the Facebook bot speakers that developed their own language, but rather to leverage what’s strange and uncanny about it.

You can better understand what intelligence is if you start harnessing absurdity. You can better understand what it is today with that, and just with paradigms that are sixty years old.

We're suggesting when we're landing with things like deep learning that we actually don't understand where it's going or where it might go. Maybe there is something really evil and terrible that happens, and the human race is going to get wiped out.

But I would say the human race is probably doing a pretty good job in trying to do that itself.

4

Exploring Science through Science Fiction

Barry Luokkala

*Barry Luokkala is Teaching Professor IInd Director of Undergraduate Physics Laboratories in the Department of Physics. He also serves as editor of the physics alumni newsletter, INTER*ACTIONS, and as curator of the Victor Bearg Physics Museum. Luokkala is the Director of the Pennsylvania Governor's School for the Sciences (PGSS) and has been involved in numerous outreach activities to promote science education.*

As a member of the faculty at Carnegie Mellon, Luokkala helped shape the introductory experimental physics course into its present form, and has created a new laboratory course for students in the pre-health professions. He also created two courses on the subject of science and science fiction: A mini-course for first-year students in the Mellon College of Science, and a full-semester course open to anyone on campus.

He has been invited to speak on several occasions on various aspects of science fiction. He has been instrumental in the design of new undergraduate science laboratories. His recently published book,Exploring Science Through Science Fiction (Springer 2014), makes science accessible to students in the nontechnical disciplines and at the same time provides enough scientific detail to also be interesting to more technically-oriented students.

The full course that I teach in the spring semester, Science and Science Fiction, is open to anybody on campus regardless of major or year. The intent of that course is very specifically to make science accessible to non-science majors and, in particular, non-technical majors.

It turns out that it's very popular among computer science majors, so half my enrollment are computer science. The other half are all over the university — art majors, drama majors, English majors, business majors, everybody — the exact audience that I was hoping to reach to expose them to science ideas in a nonthreatening kind of a way that would make it fun to learn the science rather than tedious and stuff like that.

The course progresses from the totally objective stuff like, "What is the nature of space and time? What's the universe made of?" Those purely objective things that you can measure.

We progressively get more and more subjective and personal, going through questions like, "Are we the only intelligent life form in the universe? Can we create a machine that becomes aware of its own existence?"

The next major question in the course is, "What does it mean to be human?" Cloning is a really good example of a biotechnology that might be used or abused. We try to get the students to think about using technology responsibly, thinking through the implications. That's not always possible to do. You can't imagine all possible scenarios that stem from a discovery or a technological development, but just to get them to think about stuff beyond the purely objective measurement type things.

I do that through the focus of visual sci-fi. The truth is I am a pathetically slow reader, so I don't read much. I devoured Asimov when I was in college. *The Foundation* Trilogy, I think, are among the best literary works of science fiction ever written. I also enjoyed C.S. Lewis, *The Space Trilogy*, which is really quite well done and gets into humanness and things like that.

No, most of my sci-fi knowledge is visual stuff. For instance, there are many examples in *Star Trek* where they do time travel: "The City on the Edge of Forever." That is generally recognized as one of the best episodes from the original series.

That's one of the examples that I use to talk about mechanisms for time travel, even though it's not really explained in detail what the mechanism is. It's just the big time portal. They can jump through it, and go back into the past, and mess with the past. We talk about the philosophical implications, causality, and stuff like that. Can you go back in the past and do something to change history that would result in you not existing in the future?

That's one example.

Responsible Science through Fiction

Early on in the class, I introduce Frankenstein. People know Frankenstein. Many people think of Frankenstein as the first intentional work of science fiction.

I recognize it as one of the earliest works of intentional fictional stuff that deals with the consequences of irresponsible use of science and technology, but I would go back even further than that to the 1600s. Johannes Kepler wrote a book called *The Dream* — it was written in Latin, of course — which is kind of curious.

It's a work of fiction, but it was written in Latin, which meant that he intended for it to be taken seriously by the intellectual community. It's all about imagining what the rest of the solar system would look like from the point of view of a place that is not the Earth.

Being transported by some combination of mysterious and technological means to another body — not necessarily the moon or the earth but just some other place — what would everything look like from that other perspective?

That's the one that I would identify as the first literary work of science fiction, long before the genre was ever conceived.

I include Frankenstein at the beginning of the course because we do talk about the history of science fiction in literature and on film and then on television. In the course itself, for the material, it is really focusing on Dr. Frankenstein as someone who was absolutely obsessed with his work, which many practicing scientists are.

He does this thing because he wants to do something that nobody has ever done before. That's the motivation of a lot of real science — doing something completely new that's never been tried, totally out of the box. Who digs up bodies from the grave, and cuts up their pieces, and sews them together?

He was trying to do something that nobody had ever done before, trying to endow life to something that was lifeless. In effect, it's a God complex. He wanted to be God. That isn't said in so many words in the novel that Mary Shelley wrote, but it is said on the screen in the 1931 movie: "In the name of God, now I know what it feels like to be God."

That piece of his monologue was edited out after the Hays Code[1] started to be enforced strictly, because you couldn't take the name of God in vain like that. They just blurred that out with all of the thunder and noise in the laboratory and outside the castle in the later releases of that movie.

I didn't know that he said that until I got the DVD, which had the restored original soundtrack and dialogue. I thought, "Oh, wow. That's what it's all about. Frankenstein wanted to be God. He wanted to create life out of lifelessness."

In the course, I tend to shy away from the sci-fi stuff that is basically an action movie or a video game put on the big screen. Those are the most boring to me. I want a good story. Not just: What does it take to get to the 27th level of this impossible trap that I'm in?

My wife and I just watched *The Circle* a couple of weeks ago. It is a really interesting, thought-provoking movie that makes you ask these questions: Do I want this much technology in my life? What kind of social interactions do I really want?"

I come from the generation that wasn't raised with smart phones in their pockets all the time, so I don't feel the need to have all this constant connectedness. And what happens if society moves in the direction that the official corporate policy of The Circle is. Someone asks the Emma Watson character: "Why would you want to go kayaking by yourself when there are other employees who like to go kayaking? Why don't you want to do stuff with us?"

Well, I'm the sort of person who likes to be alone occasionally, and wants to take time to be with my own thoughts. I would not want to live in a world in

[1] Named for Will H. Hays, then president of the Motion Picture Producers and Distributors of America, this code outlined what was acceptable in movies from 1930-1968. The first part of the code outline general principles of films, e.g. characters acting in acceptable and moral ways, while the second part of the code detailed specific instances that couldn't be shown, e.g. no interracial relationships were acceptable.

which Big Brother is always watching you and your coworkers want to know what you are doing at all times. That's creepy.

Whether the students who have grown up with constant connectedness, and social media, and everything, whether they think it's creepy or will that become normative, I find those things a little bit frightening.

(People who don't connect will be) the weirdo who wants to be alone every now and then. What are you hiding?

The Science in Fiction

We deal in the course with seven major questions that come up in science fiction over and over again. I hope that they will take away something about each of those seven major questions. For example, what is the nature of space and time?

Our perception of space and time has evolved since the Greeks and Isaac Newton's perception of space as a separate concept completely independent of time, to Einstein's perception that space and time are a single four-dimensional fabric and space is warped in the presence of a large mass, and stuff like that, to maybe the next way of looking at space and time as a quantum thing.

Can we unify the theory of gravitation with quantum mechanics? If they take away anything at all from that section — What is the nature of space and time? — I want them to recognize that Newton's Laws of Motion from the turn of the 17th to the 18th century, that Newton's Laws of Motion are perfectly good for everyday life.

If you want to understand how the GPS system works, you need Einstein's relativity — special relativity and general relativity — or the system just doesn't work. The modern concepts of space and time do have practical applications with modern technology.

When we talk about the properties of materials, I want them to understand a little bit about what do we understand about particle physics, and what crazy new materials can we make? Is there such a thing as transparent aluminum? Yes, there is.

Where does all this new technology come from? How does it change the way that we do things?

There's a variety of small questions embedded within the big questions. I can't really answer definitively what does it mean to be human, but if I can get the class to think a little more deeply about what are you as a human? How do you create consciousness? Can we create a machine that becomes aware of its own existence? If we do, what are our ethical responsibilities toward it?

Well, it's not Frankenstein, because Frankenstein created life from non-life, and then he abandoned it immediately. As soon as it came to life he couldn't handle what he had done. He ran out of the room. There was an abandoned child, basically, that had to learn how to be a living, conscious thing totally on its own without any direct personal relationship to guide it.

That's abandonment of one of our human responsibilities.

Separating Fact from Fiction…

When I put up this big question, "Can a machine ever become conscious?" some of them will put up their hands immediately and say, "How do you even define consciousness?" That's the point. Do we really understand enough about what consciousness is to consciously create artificial consciousness?

Have we learned anything from Frankenstein?

I would say some segments of the population certainly have and they are the voices who are saying, "Be careful what you do. You may not know what's going to happen after you do this." Scientists are driven by doing stuff that's never been done before. Then they ask the question, "Well, should I have done this?" after it's too late.

I think it's part of a human thing to have a little bit of fear of the unknown. That's a survival mechanism. If you charge blindly into an unknown situation, you could get killed. I think part of human nature is to be a little cautious, but a lot of really brilliant discoveries are made by the people who are willing to take a little risk.

But we're not going to have a Skynet from *The Terminator*. But Darth Vader could happen. The lightsaber is not going to happen and moving things with the power of your mind is not going to happen. But cybernetic humans: oh yeah.

There are segments of society who fear science and technology. They think they are the cause of everything that's evil in the world. Then there are segments of society who embrace every new technology that comes down the pipe. They wait in endless queues to buy the latest iPhone because they have to have the latest technology and everything in between.

This particular course really is intended to address some serious science questions, some fundamental science, and get non-technical people to learn some real science. Then the other issues — the philosophical and ethical issues — are the hook to get them to think about (these ideas). I warned them at the start. This course may change the way that you watch science fiction movies and it might ruin it for you.

I want them to think critically about the stuff that they're watching, but still enjoy the entertainment for what it is.

...and What's Next

When I got to the end of the manuscript for my book *Exploring Science through Science Fiction* and sent it off to the publisher, I realized, "Oh, all of these things that I wrote in the things to come section at the end of the book, they're already starting to happen." I may have to write a second edition to predict new things that haven't yet happened.

Prosthetic devices controlled by your mind. That's here already. That was just on the verge of infancy where there were baby steps being taken to how can you control a mechanical device by interfacing with the person.

That's rudimentary stuff now. You actually have the mechanical limbs from the movie *A.I. Artificial Intelligence* where the boy who was paralyzed suddenly is able to walk. He comes back to consciousness and he's got these leg braces that are interfaced to his central nervous system and he can walk with these mechanical leg braces. That's here.

I think that's just amazing that we've reached that level of capability of getting your brain to control mechanical stuff. You don't need to understand what consciousness is in order to do motor skills.

We don't have the kind of massive processing that IBM's Watson has walking around with robots, but we do have mobile robots that can do unbelievable stuff. There are biped robots that look like machine humans that can open doors, and walk outdoors, and walk up steps. They aren't tethered to a superstructure on the roof of the laboratory. They are fully mobile, incredibly dexterous. You can push them and they won't fall over.

When I see them standing up, I think to myself, "That is a massive engineering feat that they have come up with." Which is amazing. If you could connect it to the cloud, you imagine in some way there would be a way to connect that to an interconnected device.

The book was also written before the discovery of gravitational waves. Now we have even more proof that Einstein was right about what can enormously massive things do to the fabric of spacetime. Well, it's a long step from there to warp drive, but the concept of ripples in the fabric of spacetime. We now know that it's real.

There have been serious theoretical physicists who have written papers saying that it's possible in principle, but the technology to pull it off is beyond our comprehension.

But Warp drive: no, not there yet. And it's not coming soon.

Rapid DNA sequencing. We're not quite to the level of "Gattaca" where you can take a drop of blood from a baby's foot and instantly know what are all the diseases that it's likely to have, and when will it die. But DNA sequencing has gone from a process that took thirteen years to a process that now can be done in about a day. You can map your entire genome in a day.

There are potential risks for sure. Or you can begin to cure insidious diseases that previously had no cure because you could alter the genetic information and create a customized, personal cure that would work for you that might not work for somebody else.

But the big question is: Would we in the process create some super bug that would kill the entire human race?

If your genetic information is readily available to anybody in some database, how will that influence things like health insurance. There are surely evil minds out there that are contemplating those questions.

About the ETC Press

The ETC Press was founded in 2005 under the direction of Dr. Drew Davidson, the Director of Carnegie Mellon University's Entertainment Technology Center (ETC), as an academic, digital-first (but not digital only), open access publishing imprint.

What does all that mean?

The ETC Press publishes academic and trade books and singles, textbooks, academic journals, and conference proceedings that focus on issues revolving around entertainment technologies as they are applied across a variety of fields. Our authors come from a range of fields. Some are traditional academics. Some are practitioners. And some work in between. What ties them all together is their ability to write about the impact of emerging technologies and its significance in society.

In keeping with that mission, the ETC Press uses emerging technologies to design all of our books and Lulu, an on-demand publisher, to distribute our e-books and print books through all the major retail chains, such as Amazon, Barnes & Noble, Kobo, and Apple, and we work with The Game Crafter to produce tabletop games.

We don't carry an inventory ourselves. Instead, each print book is created when somebody buys a copy.

The ETC Press is also an open-access publisher, which means every book, journal, and proceeding is available as a free download. We're most interested in the sharing and spreading of ideas. We also have an agreement with the Association for Computing Machinery (ACM) to list ETC Press publications in the ACM Digital Library.

Because we're an open-access publisher, authors retain ownership of their intellectual property. We do that by releasing all of our books, journals, and proceedings under one of two Creative Commons licenses:

- **Attribution-NoDerivativeWorks-NonCommercia**l: This license allows for published works to remain intact, but versions can be created.

- **Attribution-NonCommercial-ShareAlike:** This license allows for authors to retain editorial control of their creations while also encouraging readers to collaboratively rewrite content.

This is definitely an experiment in the notion of publishing, and we invite people to participate. We are exploring what it means to “publish” across multiple media and multiple versions. We believe this is the future of publication, bridging virtual and physical media with fluid versions of publications as well as enabling the creative blurring of what constitutes reading and writing.

About dSHARP

The mission of dSHARP is to promote innovative digital research at Carnegie Mellon University via three routes:

- **Connect** scholars to collaborators and resources across the university, city, and region.
- **Research** using digital methods to ask original questions relevant to the arts, humanities, and sciences.
- **Educate** the CMU community about digital tools and research methods.

As advocates for and front-runners of digital scholarship at CMU, our work will be guided by our commitments to:

- **Equitable collaboration**: we respect the rights of all collaborators to recognition and appropriate compensation for their work.
- **Digital accessibility**: we strive to eliminate barriers to digital scholarship by creating an environment welcoming to a diverse community of practitioners and eliminating barriers to accessing the digital products we create, acquire, or maintain.
- **Sustainable research**: we value the continued access to and sustainability of digital research processes and outputs.