AI, Personhood, and Religious Humanism*

L. Karl Branting

16 July 2017

Among the oldest and most cherished dreams of humankind is the creation of an artificial person, from the Golem of ancient Jewish tradition, to Talos, the giant bronze automaton of Greek legend that guarded Crete until it was inactivated by the Argonauts, to Olympia, the mechanical ballerina in "Tales of Hoffman," to name but 3. This dream is reflected in contemporary science fiction films that often focus on the hazards of producing an artificial person whose passions and powers might match or even exceed, our own. R2D2 and C3PO are benevolent, but Ava in Ex Machina combines a film-noir dame's taboo irresistibility with superhuman guile. HAL 9000 is a disembodied psychopath, and Skynet pure murderous malevolence. We long to create artificial persons, but fear the possible consequences.

With the rise of Artificial Intelligence, AI, some think that the dream of artificial personhood may finally be within our grasp. However, some influential thinkers think that the dangers of AI must be taken very seriously. According to Stephen Hawking, for example:

> [T]he development of full artificial intelligence could spell the end of the human race.... It would take off on its own, and re-design itself at an ever increasing rate. Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded.

According to this view, AI is a Promethean technology, more powerful than anything humans have had before, but filled with peril for those unleash it. But is this view of AI realistic?

In this talk, I will explore whether AI truly has the capacity to enable the creation of artificial persons. I will do so by telling a *human* story woven from three strands: the first, the dream of creating artificial persons; the second, the quest for machines that can calculate like people; and the third, about a handful of logicians whose seemingly useless research unexpectedly turned out to hold the key to thinking machines. When AI was born from the union of these three strands, it suddenly became possible to test our beliefs about ourselves and about what it means to be a person in a new way. I'll argue that AI systems *fail* at personhood, and that the ways they fail shine a light on what it is to be human and on the foundation that anchors our religious humanism.

Let's begin our story on the shores of Lake Geneva in the summer of 1816, where the lives of 4 glamorous, talented, but scandalous young people fatefully intersected. Lord Byron, "the most brilliant star in the dazzling world of Regency London,"[1] but also

---

1 https://en.wikipedia.org/wiki/Lord_Byron [accessed 3 June 2017]

"mad, bad and dangerous to know"[2] had fled to Switzerland after abandoning his wife and baby daughter Ada. Not long afterwards, Percy Shelley and his 18-year old girlfriend, Mary Godwin, rented a house near Byron's villa. Soon, Bryon, Shelley, Mary, and Byron's personal physician and writer John Polidori developed a literary and artistic friendship. The weather that summer was dreary and wet as a result of the titanic eruption of Mount Tambora the year before—you really can't make this stuff up—so they decided to stay indoors and read literature suited to the gloomy weather, the newly popular genre of gothic stories. At some point, they decided to write their own gothic stories. Polidori wrote a story that eventually became the first published vampire story in English, the forerunner of Dracula. Shelley wrote several ghost stories, but Mary was stuck, unable to come up with anything. Then, one night she had a waking dream of a "pale student of unhallowed arts" recoiling from the stirrings of a hideous creation that he had brought to life. This story was published 2 years later as "Frankenstein, or a Modern Prometheus." Mary's work depicted not only the terrible responsibilities and dangers of creating an artificial person, but also showed that doing so would require some kind of *advanced technology*. The book alludes to the science of the day, such as Humphrey Davies' chemistry and Galvani's discovery of the role of electricity in muscle contractions, and discusses the processes of construction, animation, and education neglected in earlier artificial person stories. Critically, she explored the *moral status* of an artificial person: as a sentient being, the creature had a righteous sense of deserving happiness as much as any other person and rage at the injustice of being unfairly denied the basic rights of every person. Clearly, Frankenstein's creation had absorbed the humanistic values of his era along with that vivifying bolt of lightning. But the nature of life itself and the technology for actually generating artificial life would remain a mystery for generations to come. The first artificial gene capable of being replicated in E. Coli was created only in 2010.[3]

The second strand of our story begins, in a bit of under-appreciated historical irony, with the daughter, Ada, whom Bryon left behind when he fled to Switzerland. Ada's mother was determined that Ada would be the opposite of her poetic, impulsive ex-husband, so she trained the precocious Ada in music and mathematics. Ada designed a flying machine at 13 and at 17 fatefully encountered George Babbage, who had already designed and partly constructed an early calculating machine, the Difference Engine. This meeting led to a lifelong friendship and voluminous correspondence, notwithstanding Ada's marrying and having 3 children with William King and becoming Countess of Lovelace. Babbage made plans for a more general kind of calculating machine, the Analytical Engine, that could be programed with punched cards of the type invented for the Jacquard loom. In a pattern to be repeated later in our story, it was Ada who wrote the clearest description of the design and function of the proposed machine, and she saw more clearly than Babbage that this design wasn't limited to numerical

2 Ibid.

3 https://en.wikipedia.org/wiki/Synthetic_biology#cite_note-48

calculations, but to any calculation. Before her early tragic death, Ada wrote what is credited to be the first computer program, to compute Bernoulli numbers with the Analytical Engine. The program was correct, but the manufacturing capabilities of the mid-19th century were not adequate to produce machines with the precision, complexity, and scale needed to build a working Analytical Engine. Once again, the dream of human-like capability foundered on the shoals of inadequate technology.

The third thread involves a small group of mathematical philosophers who embarked on what seemed at the time to be an outlandishly arcane quest: to prove the validity of all of mathematics by reducing mathematics to logic. Now, Aristotle formalized rules of logic that were intended to distinguish sound from unsound arguments. Mathematical logicians weren't interested in this *rhetorical* logic, but rather used an approach similar to what you might remember from high school geometry: start with some axioms you think are indisputable, add some postulates specific to your problem area, then perform a series of logical operations that generate new statements. One of the most famous advocates of this enterprise was Bertrand Russell: a pacifist who was fired from Trinity College Cambridge and imprisoned for opposing Britain's entry into WWI, winner of the Nobel prize for literature, Gay Rights advocate, and socialist despite his aristocratic pedigree.

Russell and others showed that the truths of mathematics could be build up, step by step, by logical inference. A key insight was that mathematical statements could be viewed as just sequences of symbols, permitting logical inference to be described very precisely as the process rewriting old symbol sequences into new ones without any reference to ideas or thoughts. This might seem unintuitive and strange, but it opens the door to inference as a mechanical process requiring no human insight and to representing the inference processes themselves as sequences of symbols. Russell et al.'s breakthrough was an exciting development in a certain sub-sub-field of philosophy, but not exactly household news.

However, it soon dawned on certain visionaries that logic, rather than cogs and wheels, could be the raw material for finally constructing the Babbage/Lovelace thinking machine. One such visionary was Alan Turing, a British mathematician born in 1912, perhaps best known for 2 things: First, he led a team during WWII that cracked the Enigma cipher used for submarine communication by the German Wehrmacht, an achievement celebrated in inaccurate but good movies, like Enigma, and inaccurate but bad movies, like The Imitation Game. Second, despite his vital contributions to the survival of Britain, he was prosecuted for homosexuality under the same law that Bertrand Russell was working to abolish. A year after his prosecution, he died of cyanide poisoning. This is widely believed to have been suicide, but his mother always believed that it was because he didn't listen to her advice to wash his hands thoroughly after handling his electroplating chemicals. Please, listen to your mom! Regardless, he is an enduring martyr to the cause of Gay Rights.

In our story, however, Turing's key contribution was showing, in his 1936 dissertation, that a device that could follow a set of instructions for writing and erasing symbols on a strip of tape could perform any computation. His doctorate was actually for showing the *limits* of computation, but the important practical contribution was this completely general computing model, now known as the Turing Machine. A crucial property of the Turing Machine is that the instructions and data could be represented the same way, as symbols on tape.

As happens even today, it was the Pentagon that first turned Turing's theoretical model into a practical reality. Physicist John Mauchly and engineer Presper Eckert received DOD funding during WWII to build a computing device that eventually became EDVAC, arguably the first general-purpose programmable digital computer. EDVAC was completed over budget and too late for its original purpose, but visiting mathematician John von Neumann realized the revolutionary significance of its design and published a particularly clear description of this general-purpose architecture. Unfortunately, von Neumann neglected to mention that this architecture had been developed by Mauchly and Eckert, as a result of which this design, which is the basis of every modern computer from the ones in Mars Rovers to the one in your smart phone, is constructed following what is known as the "Von Neumann Architecture." To add insult to injury, Eckert and Mauchley, who were better engineers than businessmen, became insolvent, and their company was acquired by Remington Rand in 1950.

So, the von Neumann architecture united Bertrand Russel's dream of reducing computation to logic with Ada Lovelace's dream of a universal calculating machine. But what about the third strand, the dream of artificial personhood?

The term "Artificial Intelligence" was coined in 1955 at a Dartmouth symposium and quickly became identified with the view that both computers and people can act intelligently because both are systems with the ability to make and manipulate symbolic representations. For humans, these symbols are thoughts and mental images; for computers, they are binary strings. Year by year, AI system performance improved for inference, planning, game playing, language processing, vision, diagnosis, and other areas of human ability. Much of the improvement in the early days came from painstaking analysis and modeling of *human* problem-solving. For example, medical expert systems were constructed by interviewing and studying the behavior of doctors and representing this knowledge in a logical form that computers could interpret and apply. Once represented, this knowledge could be used for teaching, explanation, simulation, or diagnosis. The discipline of constructing such systems was a window into the nature of human expertise, providing both a language for expressing models of human behavior and a way of testing such a model by seeing whether they produce the same expert judgments as a human experts.

The emergence of AI had a dramatic effect on theories of the human mind. Throughout much of the 20th century, the dominant doctrine in US psychology departments was

Behaviorism, the view that human behavior can be explained without referring to thoughts or feelings, on the basis of conditioning alone. When I was in college, the foremost behavioralist psychologist, B.F. Skinner, published a best seller titled “Beyond Freedom and Dignity” arguing that society would improve only when we got rid of our ridiculous childish beliefs in free will and moral autonomy. This view is a deeply anti-humanist downer, and I had friends who switched majors from psychology to theology because they preferred to take a flying leap of faith rather than endure such a dispiriting world view.

AI was a spike through the heart of Behaviorism, not just because it claimed that thoughts and beliefs were real things consisting of symbolic representations necessary for intelligent behavior, but because it made it possible to *test* competing theories of intelligence. The only way to enable machines to solve human-like problems was to program them with goals, plans, concepts, and rules, all the things that Behaviorists refused to ascribe to people. Systems build *without* such symbolic representations tended to perform very badly, like the famously bad early translation system that translated the phrase “The spirit is willing but the flesh is weak” into Russian and then back into English as “The vodka is excellent but the meat is rotten.”

AI triggered what has been termed the “Cognitive Revolution” in which the information-processing model of human mind transformed psychology, linguistics, anthropology, and even philosophy and neuroscience.

While the model of human intelligence as symbol manipulation yielded increasing useful machines and legitimized the mind and cognition as areas of central to human intelligence and behavior, it didn't in itself answer the question whether a sufficiently faithful model of a human mind would, in fact, *be* a human mind. AI researchers have, for the most part, never been too interested in this question, focusing instead on making useful systems rather than imitating people. Science fiction writers, on the other hand, have written countless stories about how a sufficiently accurate model of a person would really be human, like the replicants in Blade Runner. How realistic is the idea what increasingly realistic symbolic reasoning systems could eventually have a real, as opposed to simulated mind? To restate the question, is the key thing that makes us human how we solve problems, or is it something else?

The question of what makes us human is particularly important to religious humanists. The First Principle of Unitarian Universalism is the inherent dignity and worth of every person. Usually, the meaning of “person” is clear: for example, it includes all races, sexes, nationalities, religions, ages, and occupations, but excludes portraits and manikins and dolls no matter how realistic. There are some harder cases: someone who has permanently lost all brain function and is artificially ventilated, for example, or a fertilized but unimplanted ovum. Animals are not persons, yet many of us feel that some animals have at least some of the moral status of humans and therefore shouldn't, for example, be subjected to unnecessary pain.

For many Humanists, the key requirement for having dignity and worth is *sentience*, that is, having actual or potential *consciousness* or *subjective experiences*. All humanistic morality rests on a commitment to the equivalence between one person's subjective experiences and another's. The Golden Rule—do unto others what we would have them do unto us—and the Utilitarian principle of choosing the action that does the greatest good to the greatest number, both make sense only if one person's pleasure and pain are fundamentally equivalent to another's.

How do we know that other people have the same experiences that we do? For one thing, we know how *we* act when we have an experience, so we assume that *others* how act in the same way when experiencing the same thing. Beyond simple observations, we are neurally hardwired for empathy, with portions of the insula and mirror neurons[4] that respond equally to actions by ourselves or others. Empathy begins in infancy, when the cries of one baby trigger "sympathetic distress" in others, and in adulthood vicarious, empathetic sharing of others' experiences is key both to our relationships with others and to the thrill that we get from the experiences of fictional characters.

When we turn from other people to animals or androids, the question of how we can ever really know another's subjective experience gets harder. Anyone who has seen "When Harry met Sally" knows that outward behavior is a fallible indicator of actual experience. To really be sure of other's internal mental states, we would need an answer to the question "Where does consciousness come from?" Clearly, it has something to do with the activity of the brain, but how can a set of biochemical activities, which a scientific observer can describe with accuracy, give rise to a subjective experience, which no scientific instrument can ever directly capture?

Philosopher and mathematician Rene Descartes made a proposal reflecting Christian belief of his day: the body is basically a *machine* that is animated by a *soul* that contains a person's consciousness, will, and moral status. The problem with this proposal, from the standpoint of subsequent philosophers, is that no one has even the slightest idea how a soul and a body could interact or even what soul-stuff could possibly be. Today, Descartes' proposal is generally viewed as less an explanation of consciousness than a substitution of one conundrum for another. In any case, many of us subscribe to a religion that doesn't include souls separate from the bodies that they animate.

The fact is that, remarkably, there is no plausible, testable theory of how consciousness arises from matter. It seems to be a puzzle as deep as "Why is there anything at all, rather than nothing?" Neuroscientists can connect brain activities to mental activities, but there aren't even any good suggestions for why some correspond to subjective experiences and others do not.

On the other hand, we *can* clearly say that our morality is founded on a commitment to the equivalence between our experiences. It's the foundation of our moral landscape, regardless of whether we can explain how consciousness arises from matter.

4 https://en.wikipedia.org/wiki/Mirror_neuron

Returning to the question of whether an AI system could ever achieve artificial personhood and become a moral agent like Frankenstein's creation, we have no grounds for supposing that a computational simulation of a brain, no matter how realistic, can ever give rise to consciousness the way an actual brain does. After all, there is nothing contradictory about something that *acts* like a conscious person but isn't. For example, sleep walkers can sometimes act in an apparently purposeful way and even reply to questions, but on waking have no recollection or other indication of having been conscious. The science-fiction writers' trope that every sufficiently realistic simulation of human information processing necessarily achieves actual consciousness seems as implausible as insisting that every sufficiently realistic simulation of the circulatory system would contain actual blood or that a sufficiently realistic simulated rainstorm could actually drench you with water.

In summary, we have no idea what it would take to create an AI system with consciousness and subjective experiences, but an AI system lacking these attributes also lacks moral standing. In the words of philosophy Susanna Goodin, a thing has "moral standing … if it could be harmed or benefitted for its own sake or on its own behalf. ...If I leave my hammer out all winter" it may be damaged, but "the hammer hasn't been harmed in its own self, its own sake, or on its own behalf because it doesn't have a self or a sake or a behalf."[5] Like hammers, AI systems are tools that can be useful if working well and useless if broken. But, lacking consciousness, they have no self, or sake, or behalf of their own.

This is not to say that all the fears of AI are unfounded. AI is a technology for automating the tasks that people perform, and as AI techniques improve, an increasing range of jobs are threatened with obsolescence. This is part of a process that has been going on for centuries. The Jacquard loom whose punched cards Ada Lovelace used for representing programs caused riots by putting weavers out of work. However, the very effectiveness of AI techniques at duplicating human performance gives them an unsurpassed capability for job displacement. Even AI researchers are uncertain what jobs will be safe for their children and grandchildren. But the fears that artificial persons will become *competitors* to the human race seem misplaced for the foreseeable future. Our future android companions may be tools that satisfy human social needs by simulating human behavior, but nothing more.

I've tried to present AI as a human story in two different ways. First are the historical personalities: Ada Lovelace, whom I like to imagine elegantly attired in her spacious London lodgings, writing the world's first computer program as servants bring her cups of tea and shoo her children and husband from the room; Bertrand Russell, puffing on his pipe in Cambridge as he worked to reduce all mathematics to logic and to replace war and religious superstition with a benevolent socialism, then later, in his 80's, as a Viet Nam War protester and author of a best-selling autobiography detailing his

5 Personal communication.

romantic and intellectual passions; Alan Turing, consoling himself for the tragic early death of the young man he loved by designing a universal computing mechanism; John von Neumann, Hungarian child prodigy and polymath, sharing off-color Yiddish jokes with his colleagues at the Princeton Institute for Advanced Studies and enduring the teasing of his wife who said that the only thing he couldn't count was calories. Some stories had sadder endings. John Polidori, having given the world the first vampire novel, became despondent over debts and killed himself at age 25, swallowing the same poison as would kill Turing a century later. Byron, ever the foe of authority figures, died fighting for Greek independence, and Shelley, lover of nature, was lost in a storm while sailing alone, leaving Mary to support their only surviving child. While Mary focused on publicizing her late husband's work, she published 5 additional novels, including what may be the world's first post-apocalyptic novel, The Last Man.

The story of AI is also a human story in that designing systems that solve problems the way we do requires understanding how we work. AI's view of intelligence as the ability to reason symbolically triggered a cognitive revolution that vanquished the stifling reductionism that dominated psychology for much of the 20th century, and the limitations of this model, reflected in the current surge of interest in neural networks and deep learning, suggest that our instinctive and emotional side is a central to our humanness as logic and deliberation.

Finally, pondering whether AI systems have the potential for personhood can help focus our thinking on the basis of our humanistic morality. Every human has dignity and worth because we all share the unique gift of consciousness and the same spectrum of feelings, and are bound to one another through threads of empathy and compassion. Our greatest fear about AI shouldn't be that it may become our rival, but that it will be a tool that we use for mischief in an all-too-human way. Our perception of even the simplest robot as animate and sensate says much more about us than about them: we are inherently social beings whose sensitivity to even the most subtle social cues, evolved over thousands of generations, leaves us helpless to resist anthropomorphizing the machines we have created. The empathy that makes us vulnerable to this illusion is basis of our morality and the source of our compassion.

I conclude with the hope that we may have the wisdom to use each new technology in creative and constructive ways. And may our awareness of our shared humanity always guide us to treat all actual people with the compassion and generosity that their inherent dignity and worth makes them deserve.