Peter Mountanos

Professor Dennis Shasha

Gallatin School of Individualized Study

1 March 2016

Rationale: Humanity’s Quest for Artificial Intelligence

*Introduction*

Humanity’s desire to replicate its most valuable asset—intelligence—emerged in antiquity, and is today, more than ever, at the forefront of research. Our written fascination with automata and artificial intelligence (AI) dates back to Ancient Greece, where *The Iliad* describes golden women built by Hephaestus who “were like real young women, with sense and reason, voice also and strength, and all the learning of the immortals” (*Iliad* 18.417-421). For millennia, humanity expressed this desire mainly through art forms, due to limitations in technical understanding and abilities. However, in the mid-twentieth century humanity began to make substantial technical progress in the field of artificial intelligence. Despite this progress, our ability to build a machine with intelligence indistinguishable from ours is still highly debated. Through my colloquium, I hope to gain insight into the potential of machine intelligence, and the consequences we may face if we reach the point of singularity—a time in which artificial intelligence will overtake human thought.

While evidence for the potential of the singularity may be inconclusive, there is little question that we are only at the dawn of artificial intelligence. Every day, we interact with primitive forms of machine intelligence, ranging from querying our virtual assistances on our smartphones to being recommended products on the web, which are algorithmically inferred from our past preferences. It is difficult to ascertain exactly how these systems will continue to develop and improve, thus making it problematic to predict when this day of singularity will come—if it ever does. Ray Kurzweil, in *The Singularity is Near*, suggests that this date is 2045, yet other timelines proposed in his book have fallen short [or have not come to pass]. Regardless, developments in machine learning are occurring at an unprecedented rate, which suggests that it is imperative to consider both the benefits and the ramifications of these innovations on human life.

*Concentration vis-à-vis Colloquium*

Over the last three and half years, I have studied the deeply intertwined nature of computer science and mathematics. In an applied sense, I combine these two fields to program computers to extract knowledge from, and to learn from, data. In today’s terms, this field is generally called data science. However, as it is in its early stages, the term data science is not well defined. Some see it as a unique combination of fields such as machine learning, statistics, predictive analytics, artificial intelligence, and the like; others argue that it is simply a rebranding of well-established methods. In either case, data science is an extremely interdisciplinary field of study. Its research has implications beyond computer science and mathematics alone, as these are simply the tools employed to solve some of the most challenging problems in our world today. In fact, data science is used to solve problems in natural language processing, genomics, finance, public policy, healthcare, and even high-energy physics.

Throughout my time at Gallatin, I have focused my efforts on understanding the theory behind machine learning, as well as applying this theory to real-world problems. Specifically, my main area of research lies in the field of recommender systems—pieces of software that aim to recommend products or information to users based on certain preferences. Along with Panagiotis Adamopoulos, of the Stern School of Business, I have worked to create a model of recommendations combining users’ past preferences (a standard approach in the field, commonly referred to as collaborative filtering) with features of their personalities, in order to improve recommendation accuracy and diversity. This research has included work beyond just machine learning, such as understanding human motives from a psychological perspective. While my research interests are much too applied to be the focus of my colloquium, the notion of the possibility of artificial intelligence provides a much broader base to examine a diverse set of existing literature, both in antiquity and in modernity.

*Historical Beginnings*

In order to best understand the prospects of AI, it is important to consider its early foundations—much of which began in Ancient Greek literature and Renaissance philosophy. McCorduck, in *Machines Who Think*, claims that artificial intelligence was “an ancient wish to forge the gods.” Automata capable of movement and reason are described in ancient texts, paintings, sculptures, and plays. Homer, in his epic poem The *Iliad*, makes mention of golden tripods that had “wheels underneath the base of each one so that of their own motion they could wheel into the immortal gathering, and return to his house” (*Iliad* 18. 371). Not only were intelligent devices conceptualized by artists and writers, but also by philosophers, such as Aristotle. In his work, *Politics*, he considers the existence of automata, and what effect it may have on mankind.

For suppose that every tool we had could perform its task, either at our bidding or itself perceiving the need, and if – like the statues made by Daedalus or the tripods of Hephaestus, of which the poet says that “self-moved they enter the assembly of gods” – shuttles in a loom could fly to and fro and a plucker play a lyre of their own accord, then master craftsmen would have no need of servants nor masters of slaves. (pp. 65)

Although Aristotle considers the benefits of automata in the context of his time with masters and slaves, it is quite similar to what AI enthusiasts promote today—using AI to delegate mechanical and mundane tasks to machines so that humans can focus their efforts elsewhere, such as on the creative process. Not only did Aristotle dream of machines that can reason and perform human-like tasks, but he was also the first person in history to scientifically explore and formulate human thought. In *Prior Analytics*, Aristotle defines several different types of syllogisms, which are arguments consisting of at least two premises and a conclusion. Aristotle’s exploration of logic heavily inspired Western thought, and laid the groundwork for the field of mathematical logic—a crucial construct towards replicating human-knowledge systems.

*Human Emotion*

While early mathematics and logic played a crucial role in the technical developments of AI, understanding intelligence and the human mind from a philosophical perspective is an important consideration when debating the possibility of AI. There have been two major schools of thought concerning the human mind—the dualist view and the materialist view. The dualist view proposes that the human mind is non-mechanistic, as it is a metaphysical object, free from the laws that govern the body. Contrastingly, materialists believe that the mind can be explained in a physical sense, and thus cognitive thought and emotions can be derived through a set of laws that define the mind and body. Many proponents of artificial intelligence tend to favor the latter view, while critics prefer the former.

The dualist, René Descartes, in his *Discourse on the Method*, was one of the first philosophers to consider whether machines, or in his case animals, could think as humans do. He argues that a machine could perhaps perform human-like actions, such as responding to questions; however, these actions would not be performed in the same way that humans would. The reason is that the machine would have to operate on principles of logic and rationality, while the human mind is free of these behaviors. Thus, in his specific example, a machine would not be able to rearrange its words to answer the question in a different way, as it is bounded by a set of laws. While Descartes’ notions have some validity, his arguments don’t map back to the real world. It is much easier to think of the mind and body as separate entities in the case of one human, but it becomes much more difficult to apply these assumptions to the behavior and interaction amongst several humans, or an entire society. Still, through his thought experiment, Descartes laid the groundwork for what eventually became known as the Turing test.

The Turing test, proposed by Alan Turing in 1950, attempts to define a scenario that would be able to distinguish whether or not a machine could exhibit intelligent behavior matching, or exceeding, that of a human. However, Turing’s definition of whether machines could think fails to address emotion. Even if a machine could respond to questions well enough to trick a human into thinking it is another human, does that necessarily ensure the computer can think? Do we define thought and our intelligence solely on the ability to answer a set of questions? Surely we appreciate more, such as individuality and emotion. Thus, the question still remains—just because we can build machines that can make decisions autonomously, does that mean that these machines can have emotions, and that they can feel as humans feel? I believe the answer to that question is entirely determined by one’s definition of emotion. In what capacity does one feel? Is it state beyond our comprehension? Or, is it a series of mechanical processes that occur as a response to a stimulus? If it is the latter, then surely we will eventually have the computing capacity to teach a computer to feel, as all it requires is enough training examples to generalize a response. However, if it is the former, it becomes nearly impossible to program a computer to feel, as emotion is something beyond even our comprehension as a race. I do not believe these somewhat rhetorical questions have a right or wrong answer, yet they are important to consider before one can conclusively decide whether machines have the ability to obtain human-like intelligence.

*The Future of AI*

Due to innovations in modern machine learning, we have built machines that excel at generalizing from previously given examples. For example, if we “show” a computer millions of images, and a caption describing those images, there are algorithms that can then generate a caption for a new image. These algorithms work by loosely simulating the behavior of the neurons in our brain—a class of algorithms commonly referred to as neural networks—more recently referred to as deep learning, which are neural networks with many hidden layers. Today, not only do we have computers excelling at image recognition, but also tasks such as game playing, natural language processing (such as sentiment analysis and machine translation), speech recognition, autonomous driving, and more. These are all tasks that humans are capable of performing, and in a sense, we learn very similarly to how computers learn. The question is, once computing power becomes great enough, will it simply become a matter of machines learning emotions, similar to how they learn to distinguish images? Due to Moore’s law, we are fairly confident that we will reach a point computationally where extremely complex problems can be solved. However, it becomes much more of an issue of framing the problem in a way in which it can become solved. Does the need for a human to represent the problem to the machine inherently rule out true artificial intelligence, or will we reach a point where machines will be able to transcend humanity and recursively self-improve? And will this self-improvement be enough to become human-like?

 While it is thought-provoking to debate the feasibility of artificial intelligence from a philosophical perspective, many scientists truly believe that the human brain is a finite machine, which can, and will be, reproduced from inorganic matter. Thus, it is something we must take seriously, and we must consider the consequences of this. OpenAI, a nonprofit research group, was recently formed, with notable investors such as Elon Musk (of Tesla an SpaceX) and Peter Thiel (of PayPal) contributing funds. OpenAI’s goal is to push forth a research agenda that leads to artificial intelligence being controlled only for the benefit of humankind. While robots rising up to destroy humanity may seem like science fiction, as Noah Chomsky argues, even the potential of this should lead us to work to prevent it. We must first try to conceptualize a future with AI in it, in order to steer developments in that direction. Will we live in a world where we augment our natural cognitive abilities with superior computational power, or will it be a world where the machines that we have created become our rulers? Companies like OpenAI serve an important purpose, as we must proceed with our research endeavors with great caution.

*Conclusion*

In my concluding remarks, I will pose three questions that I hope to discuss in my colloquium:

1. Is the mind a purely mechanical object responding via interactions between neurons, or is it an independent metaphysical object, as suggested by Descartes?
2. Is it possible to create a machine that has intelligence indistinguishable from that of man’s?
3. If possible, what developments in machine learning and artificial intelligence are needed to reach this state?
4. How about the moral issues?

 I hope that through understanding both ancient and modern texts, as well as recent technical developments in AI, I will be able to glean better insight onto these questions. My current position holds that we will be able to reach a point of human-like artificial intelligence, although I am not necessarily convinced that the singularity is as near as 2045, as Kurzweil suggests.

Booklist

*Ancient, Medieval and Renaissance Classics*

1. *Iliad* - Homer
2. *Prior Analytics* - Aristotle
3. *Politics* - Aristotle
4. The Bible, King James Version
5. *Utopia* – Sir Thomas More
6. *The Prince* – Niccolò Machiavelli
7. *Discourse on the Method* – René Descartes

*Modernity—The Humanities*

1. Frankenstein – Mary Shelley
2. Machines Who Think – Pamela McCorduck
3. “The Sandman” – E.T.A Hoffman
4. R.U.R. (Rossum’s Universal Robots) – Karel Capek

*Modernity—The Social and Natural Sciences*

1. “Computing Machinery and Intelligence” – Alan Turing
2. *Cybernetics* – Norbert Wiener
3. *The Computer and the Brain* – John von Neumann
4. *Big Data* – Mayer-Schönberger & Cukier

*Area of Concentration*

1. *Fooled by Randomness* – Nassim Taleb
2. *The Signal and the Noise* – Nate Silver
3. *The Singularity is Near* – Ray Kurzweil
4. *Natural Computing* – Dennis Shasha
5. *Flash Boys* – Michael Lewis