May 24, 2016

Reflections on Artificial Intelligence

I am a computer engineering major who came into this class with a general knowledge of the state of current technology and a bit about its projected impact on society. My goal in taking the class was to build a clearer and more interdisciplinary picture of the problem space and the range of proposed solutions, fueled by my drive to be conscious of the impact of my work on the lives of other people. In the array of six statements, my initial position was most closely aligned with Statement 5 coupled with elements of Statement 3. I knew about the extraordinary potential of A.I. to automate tasks, and the argument that generalizable machine learning algorithms make this different from the industrial revolution; I had heard of machines writing music, but I was of the opinion that though computers can mimic emotion, they cannot fully feel it the way a human can.

My views as the class comes to a close can be described as still mostly in that general theme – in particular, I still most closely align with Statement 5 – but there are some differences, some additional layers of nuance and caveat, and some flavors of alternative perspectives. I am still uncertain about some things, but I have more cognitive tools now with which to approach these tough questions.

I'll first address the potential impact of A.I. on jobs. Certainly A.I. is already taking over a number of today's jobs, especially moderately skilled white-collar ones, and it threatens many others, both in the professional sectors and in manual labor; the trend has been well documented and there is nothing to suggest it won't continue. Both Humans Need Not Apply and Rise of the Robots addressed this in detail; they both confirmed my understanding of the threat and exceeded my expectations in how far the technological takeover has already come, such as agricultural robots that can identify ripe fruit and pick it. Jobs that require "intuition" are not safe: intuition is our human way of taking mental shortcuts and simplifications, and expert knowledge often boils down to forming useful cognitive models, but machine learning can find just such patterns in the data. Furthermore, though human-A.I. collaboration may be feasible in the short term, with computers employing their computational accuracy to dominate some tasks and humans using their "common sense" to sanity-check computers' work, over time computers will become ever more effective in the work they do, to the point where humans will become distinguished on the job primarily by the errors they make, at which point it will not be economically efficient for them to be employed in such a role.

I am convinced that this technological revolution is not like previous ones in human history, in terms of both the sheer pace of technological change (as deftly compared to the industrialization of the agriculture sector in Humans Need Not Apply) and the type of change (generalizable, self-improving systems that do not create new jobs as quickly as they devour them, as thoroughly evaluated in Rise of the Robots). Furthermore, I find it more than plausible that the benefits of A.I. will be concentrated largely in the hands of those who previously had the funds to invest in making the most of it, resulting in extreme income inequality even more so than we see today; the degree to which this is possible, as revealed through research and discussion, also surprised me.

The question of where the economic impact of A.I. will be felt is an interesting one, and one I hadn't thought much about before this class. Much analysis has been done specific to developed countries and especially the US. However, the assessments we did of African nations

made it clear to me that second- and third-world countries have a much rougher path ahead of them in the struggle to stay relevant in the global market. Though developed countries will certainly feel the growing pains of A.I., they at least have more power and tools to address the oncoming challenges, and I do believe they will be able to grow in the end. The future is less certain for countries with limited resources and a small tech sector.

Economic transformation is coming; some aspects of it are already here. I'd estimate that the effects on the job market will be felt keenly in ten years and will be quite painful in 25, and within 50 years the job market will bear fairly little resemblance to its state today. I'll add the disclaimer, though, that exponentiation is hard for human minds to truly grasp – including mine – and so this process may very well happen even faster.

At the point where computers are able to perform most (if not all) economically productive activities more efficiently than humans can, humans will need to better define our role as a species and our place in the world. One key area we are sure to cling to is our capacity for emotional processing and personal connection; each of the three movies we watched voiced this feeling in some way. However, this view needs a more careful examination in light of advancing technology. My perspective is this: A.I. will evolve to the point of writing poetry, music, or creative art or literature that is indistinguishable in quality from human efforts; though this sophisticated mimicry does not automatically imply that these computers experience being human, it does prompt us to refine our definition of the human experience.

I am disinclined to adopt blanket statements about computers someday becoming indistinguishable from humans modulo extraordinary mental powers, at least at this point in time. I'm putting aside here the caveat of physical distinction: if it's a non-humanoid and/or distributed system, it's not human, and if it bleeds, it's not a computer. (This in itself gets into questions of bionic enhancement that are a bit beyond the scope of this discussion.) However, I do think computers and humans will become indistinguishable in certain ways. I'll explain my perspective in three slices of distinction.

First, I'll address the definition of "experience of the world." Consider the thought experiment of Mary in the black and white room as mentioned in Ex Machina. If Mary really knows all there is to know about color, save the actual experience of it, what she gains by leaving the black and white room is the coherent chronological constant stream of visual information arriving through visual receptors she hadn't previously used, then filtered and simplified in order to make sense of the world around her. If this is how we define perception as a human experiences it, then it does not seem too much of a stretch to my imagination that we should be able to engineer this some day for a computer. The key challenge will come in determining how to filter in ways that draw out the important semantic information in a scene, but I don't think this is insurmountable, based on current progress in machine perception.

Second, I'll consider the distinction between mimicking emotion and feeling it. Although I disagree with some of the points Julie Villegas put forth to the class, here I echo her concern about emotional authenticity. As a human example, actors are trained to exhibit emotions of characters they are not: they can display with extraordinary precision extreme pain, enlightenment, or mania without actually being or having been pained, enlightened, or maniacal. There is an element of "getting into character," but ultimately the displayed emotion is monitored by a quiet sanity check that reminds the actor not to actually strangle the target of feigned fury. Computers, likewise, may display emotion in seemingly human ways without

actually feeling those emotions. For the most part, the difference will not be substantive. But I have yet to encounter evidence that computers working under the utility functions of efficiency and performance can adopt truly self-inhibiting behaviors beyond their control as a result of these emotions. Namely, can computers grieve? Can computers struggle with identity and conformity as many American teenagers do? Can computers commit suicide? It serves no purpose in their design. I am not suggesting that these behaviors cannot be mimicked if programmed in, nor am I suggesting that it is impossible for such behaviors to emerge sometime in the far future. But there is still something about the human condition that we don't understand well enough to translate into machine terms.

Third, I'll turn this around and consider the pitfalls of taking this approach without first clearly defining the human experience. If we restrict our definition of humanity in terms of what true emotion is, what human cognitive or metacognitive function is, and so on, then are we to say that autistic people, people with Down syndrome, or sociopaths are somehow less human? Certainly my neighbors with an autistic son would argue against that. In this vein we need to consider our metrics of humanity carefully and be willing to gerrymander our definitions or accept the logical consequences.

At least in the relatively near term, there will still be a significant difference between human cognition and computer processing. But I think eventually this distinction will become vanishingly thin, starting with our inability to say with certainty "ah, yes, that was done by a human." Yes, computers will have extraordinary mental powers, including those of simulating the output of a human mind. Will they pose an existential threat to humanity? No, in my opinion. We will survive as a species. The main question is what our existence will be like in an A.I. world.

What does all this mean for me as a computer engineering major? Computing still has the power to make a positive impact on people's lives, but there are serious risks that should be addressed as technology continues to progress. Computing solutions must be developed with humans in mind, especially with an eye toward employment opportunities, potential for widespread economic advancement, and mental health consequences. My current focus is in computer security, which I anticipate will become even more critical as our dependence on our technological infrastructure increases and we rely ever more on computer systems to make decisions. The field of security is also difficult enough to automate that I'll be able to make a positive contribution to it in the foreseeable future. Beyond that, I'll be more conscious of the influence of A.I. on current events and the socioeconomic ramifications of A.I. advancements I read about, and more prepared to communicate my thoughts.