Greg Yeutter
January 16, 2015
CS510 HW1
Part 1: Written Problems

1.A: Musk
Musk and others, including Stephen Hawking and Robert Bishop, warn that we must be careful with the application of AI in fear of a technological singularity that puts humanity at a serious disadvantage. Musk is not very specific in his warning, other than that we should proceed with caution, but Hawking worries that robots may soon outsmart humans and continue to progress at a rate faster than human evolution. Bishop disagrees that we can exactly recreate human intelligence in robots any time soon, since the definition of intelligence is unclear and changing, but warns that decision-making functions of military equipment could lead to terrifying consequences, even if the intelligence employed by these devices is nowhere near the superhuman level [1].

Within the past week, an open letter appeared online pushing for robust and beneficial artificial intelligence, signed by several hundred individuals including Musk, Hawking, and Peter Norvig. The letter posits that the field is growing rapidly, even exponentially, and the potential benefits of AI for society will also grow at an exponentially in the coming years. It warns, however, that this growth of benefits could also lead to a growth of pitfalls, so future research should be driven toward the benefits of AI-associated technologies [2].

Indeed, we now have cars that can drive themselves, potentially better than humans can, which would provide enormous benefits to the safety and productivity of society (at least those with high human capital). However, the ethics of robotic cars are tricky: is it better to hit and kill a pedestrian vs. swerving off the road, killing the vehicle occupant in an unavoidable accident? Statistically, AI should reduce these types of scenarios, but when they do inevitably occur, what is the best course of action, and who is liable? Because we are now thinking about these ethical dilemmas, it shows that we are pushing toward AI that is beneficial to society. But we must make sure, as others have warned, that we stay on this path. As technologies continue to advance, more than an ethical code of conduct may be required to ensure that AI development remains beneficial.

1.B: Chinese Room
The Chinese room argument refutes the idea that a computer, given a set of inputs and rules for processing them, thus providing an output that is identical to what a human would output in the same scenario, thereby exhibits the same consciousness as humans. In other words, Searle does not agree that the mind is just a computer program, and there is a difference between processing a task based on rules and actually having a conscious understanding of the information.

He uses two examples to make his point. The first is the Chinese room, in which he is fed Chinese characters, and must provide an output from these characters based on a rule book. Since Searle does not understand Chinese, he is merely acting as a computer program that follows rules without understanding what is actually going on. He contrasts this with receiving

the same question in English, where he can give a seemingly identical answer that is based on more than rules of the language but a conscious understanding of the question.

The second example is a model of human digestion. The computer can model all the steps of human digestion with as much detail as the authors specify, but the computer cannot physically digest a pizza with the same mechanism. Thus, the computer program is simply a model, not a replica, of the human digestion process.

In terms of understanding, human consciousness is obviously more than a set of hard-coded rules. We do not simply know English from birth; instead, we must learn it over time through contact, association, trial and error, and other techniques. An English speaker could learn the syntax rules of another language, such as Japanese, and successfully output a response without ever understanding the context. But learning to get an understanding of a material is a structured, not factored, process. We apply our knowledge by synthesizing material explicitly, not by having a Boolean representation of an ultra-specific scenario such as *TruckAheadBackingIntoDairy-FarmDrivewayBlockedByLooseCow.* In other words, we can navigate unknown scenarios because of associations and context.

In the same way, AI is about more than hard-coded rules: associations through various learning techniques can lead to something that approaches what we perceive as consciousness. It is unclear whether we will ever fully understand consciousness, and by extension be able to implement computer programs that are identical, but it is possible that given the appropriate agent, sensors, and actuators, we can build machines that do more than follow rules but intelligently make decisions based on an understanding of context.

1.C.a: Small Towers of Hanoi

- Performance: Completion (successfully moving entire stack following all rules), Efficiency (best solution = smallest number of moves)(for 3 disks, smallest number of moves is 5).
- Environment: Three rods and three disks.
- Actions: (For each disk D_i on each rod R_i) move disk up (pop D_i from R_i) + move disk down (append D_i to R_i), counted as 1 move.
- Sensors: Each rod R_i perceived as a stack/list (last in, first out). Length of each rod stack R_i can be 0 (no disks) to 3 (all disks). Size of each disk based on its number (D_1 = smallest, D_3 = largest). Attempting to append a larger D_i than the value currently at the end of the rod stack (e.g. appending D_3 to rod R_i containing D_1): move rejected. Any disk D_i cannot be on more than one rod R_i simultaneously.

1.C.b: Pac Man

-Performance: Completion when no F (food/pellets) exist within the map.

-Environment: Map with pellets and walls.

-Actions: P (Pac Man) moves up, down, left, or right to adjacent cell. Cannot move diagonally (valid move is (x_1, y_1) to (x_1, y_2) for example, but not (x_1, y_1) to (x_2, y_2)).

-Sensors: Map represented as an *n*x*m* matrix. Each position in the matrix may have the char value E (empty), P (Pac Man), W (wall), or F (food/pellet). If P enters a cell with F, F deleted and replaced by P. If P leaves a cell, it has a value of E. If P attempts to enter a cell with W, the move is rejected. P can freely enter a cell with E (E is deleted upon P entering, then reapplied when P leaves).

References

[1] http://www.independent.co.uk/news/science/stephen-hawking-right-about-dangers-of-ai-but-for-the-wrong-reasons-says-eminent-computer-expert-9908450.html

[2] http://futureoflife.org/misc/open_letter