Connor Johnson

FP 0006

2/12/2018

Controlling Artificial Intelligence

Over the past century, computers have gone through huge innovation and been designed to do some amazing things. As scientists and engineers continue to develop computers, we will get closer to achieving a humanlike artificial intelligence. A large problem with AI like this is control. By creating a machine that is as intelligent as we are, we create the potential threat of losing control over what we created. There is no point in building something that we don’t have control over. It is important to understand that AI is not simply a machine like a car where we understand everything that is going on. An AI is much more complicated and “human” than any other machine. It is not always known how an AI reaches the conclusion it does similar to a human. Because of this, it is important that we design AI cautiously so that we can stay in control of it. This idea of controlling AI can be seen in Isaac Asimov’s short story, *Little Lost Robot.*

At its basics, artificial intelligence is a decision-making machine. The purpose of an AI is to take in information either from a user or from the world around it and use that information to decide which option will result in the best outcome. This can be done using minimax search algorithms or Monte Carlo tree searches for games or neural networks for other uses. The most important thing to note about these algorithms is that the programmer doesn’t always know what is going on inside them. Thus, a robot can have “thoughts” that the programmer never intended to occur. For example, in Isaac Asimov’s story, Dr. Susan Calvin says, “Granted that robot must follow orders, but subconsciously, there is resentment. It will become more important than ever for it to prove that it is superior.” This kind of thing is impossible to avoid when designing AI, but it is possible to control it. It is impossible to control the actual thoughts of a robot, but it is possible to control how these thoughts are used to make decisions. This can be done using rules that prevent a robot from doing certain actions. This can be seen in *Little Lost Robot*. Each robot has three rules in its system that it can’t break: don’t injure a human through action or inaction, obey a human’s command unless it conflicts with the first rule, and protect your own existence unless it conflicts with the first two rules. These rules are very important as they determine how a robot will act.

Since the rules of the robot are such a vital part of how it acts, it is important that the rules are chosen carefully. Then, they must be extensively tested to assure that they work properly. It can be very consequential if this is not done correctly. Asimov shows this in *Little Lost Robot*. In the story, some workers were doing work that involved gamma rays. In order to do their work, they had to expose themselves to gamma radiation that was completely harmless. The robot would sense the radiation and see it as a threat. The robot would then rush in and pull the worker from what he was doing to keep him safe. To keep this form happening, they reprogrammed the first rule so that the robot could allow a human to be harmed through inaction. The robot still wasn’t allowed to harm a human. Dr. Calvin explains why this could be potentially dangerous. She explains that a robot dropping a heavy weight on a person wouldn’t be breaking the modified first law. She said, “once the weight left his fingers, he would be no longer the active medium. Only the blind force of gravity would be that. The robot could then change his mind and merely by inaction, allow the weight to strike.” This simple change in the rule would allow a robot to kill a man, even though according to his protocols, he is not allowed to. The rules of a robot can be thought of as the reflexes of a human being. When a human touches a hot stove, he/she will pull their hand away before any pain is felt. The spinal cord realizes the stove is hot before the brain does and reacts accordingly. Your conscious mind never made a decision. Similar to humans, robots have rules that affect decision making and will try to get around them. Since AI is very good at working with logical problems. It will often find a workaround to its rules. So, an AI must be programmed carefully to block out all bad decision making.

While a robot is very similar to a human being in the way we think, there are many differences in the way we communicate. Communication between two robots is simply a passage of information. Data is simply transferred from one robot to the other. It’s like reading from a textbook. There’s nothing happening other than information transfer. On the other hand, communication between two humans is much different. A conversation between people involves a transfer of information as well as emotions, facial expressions, figures of speech, and much more. There’s a reason people prefer face to face conversation over texting. A conversation between two people involves so much more than an exchange of information. Because of this, exchanging information between humans and robots can be very difficult. To this day, there doesn’t exist a robot that understands sarcasm and most don’t understand what a figure of speech is. So, there can be some confusion between humans and robots. This can cause some problems especially seen in *Little Lost Robot*. In Asimov’s story, they were trying to find a robot that had hidden itself among a group of other robots. The robot was trying to hide from all the humans after misunderstanding a command that it had received. One of the workers had gotten mad at the robot for bothering him and said “Lose yourself” to it. Obviously, this is a common saying that many people use which means to “go away.” The robot took it literally and determined that it had to keep hidden from all humans in the station. This robot then went as far as attempting to kill someone to keep itself hidden. Something like this happening would be very problematic because it would be much more difficult to control the robot. Its responses become very unpredictable which isn’t good for the user. One solution to this problem would be to create a strict list of commands that the user is allowed to use. These commands would be extensively tested to make sure they don’t cause any problems. The problem with this is what if the user doesn’t follow this set. They could give the AI an incorrect command that results in an incorrect response. Because the user of computers usually can’t be considered trustworthy, this solution becomes very problematic. The only other way to fix this is to program a computer to communicate like a human does. Human communication is way more complicated than robot communication. This solution would be very difficult as it is very difficult to mimic how humans communicate. A robot must understand when a human is sarcastic or serious, talking figuratively or literally, and the emotions a human has when speaking. This is the only way the AI can then truly understand a human’s commands. Since this is very difficult to design an implement, it would mean that the timeline for development of AI must be extended, but it is the only way to keep AI safe when it is used in everyday life. Otherwise, the response of AI would be very unpredictable.

A long term goal if AI and robots is to implement them into everyday life. Having robots handle certain jobs instead of humans would be much more efficient and safe. Since robots are not like other machines, this must be done carefully. We’ve all seen Hollywood movies that show a robot revolution resulting in a world controlled by AI. This type of situation is realistic, but it is very possible for us to lose control over the robots we create which could be very problematic. Thus, we must be very careful in the way robots are programmed and used to always maintain control over what we create. Although it is much easier to create an efficient robot than a safe one, we must put safety first. It is the only way that humans would be able to live in a world that is filled with artificial intelligence