|  |
| --- |
| Artificial General Intelligence and The Great Tea Making Crisis |
| Ethics Blog Post |
| \*insert exam number here\* |

# Introduction:

As we continue in our research and development of Artificial Intelligence we move forward towards one of our ultimate goals, Artificial General Intelligence (AGI).

Forbes defines an Artificial General Intelligence as a machine that can [1]“perform any intellectual task a human being can”.

This means that an AGI would be capable of attempting any action without being specifically programmed to do so. It might fail, but an AGI would also be able to learn from this through trial and error. This means that an AGI would not need to be taught how to do a task, much like how a baby isn’t taught how to walk. It learns from its mistakes and improves its next attempt.

However, there is concern about these types of machines, and more specifically, what sort of unintended results might occur when allowing one of these AGIs to run. If these concerns were to be valid (and indeed some are) then there could be implications concerning the safety of data, the internet, the physical world and maybe even human lives themselves.

That may seem a tad dramatic, but in the main body of this post I will illustrate just how badly things can go wrong as well as some examples of where these worst-case scenarios may take place.

Because of these concerns there are some ethical dilemmas involved. As these AGIs could have such severe consequences should they go wrong, should we even attempt to build them in the first place? If we are ever able to, should we regulate who has access to them? Improper use or setup could lead to bad and/or unexpected outcomes. Also, some users may create and use AGIs for illegal applications.

In this blog post we address the potential ethical concerns, as well as methods for preventing bad and unexpected outcomes.

# Defining a task:

Imagine we have this robot in our lab who is programmed with an AGI but does not yet have a task to perform. We will be giving our robot the task of making us tea. As simple as this may seem, many, many problems can arise from a badly designed task.

We will also be introducing a new term here called the AGIs Utility Function (UF). A Utility Function [2] “Quantifies how we “value” outcomes”, in our case, if the robot has made us tea or not. The robot will try to maximise this utility function. What this means is that it gets some points (for example 10 points) for making a cup of tea, and it will try to get as many points as possible.

So, we want to give our robot the task of making tea. This might seem simple, but we can run into many problems. When defining a task, we must make sure that the robot wants what we want, otherwise it will do something that we don’t want it to do. This might seem like a simple concept, but it is very important. Computers take things incredibly literally, so if we tell our robot to make us tea, it will!

You would think that it would make us a cup of tea when we want tea, but this is not the case. It is trying to maximise its UF, or simply put, it is trying to score as many points as possible. It will immediately make us a cup of tea, give it to us, and then repeat this infinitely, so that it keeps scoring more points.

This is a problem as the robot will be constantly producing tea for us. However, as this is an AGI, it will be able to figure out that there is a better way to get points. It might use your credit card to pay some service to ship you thousands of cups of tea, or you might come back one day and find that your kitchen has been converted into a tea producing factory!

This is what I meant when I said we need to make sure that the robot wants what we want. If, when we want a cup of tea, we tell the robot to make tea, it wont take into account if we want it or not, it will just make as much tea as possible.

You might be thinking, “ok, well just tell the robot to make us tea when we want it to”. That would seem like the next logical step, but that won’t work either.

It will begin to make tea when we want it to, but then it will try to figure out a better way to score points. It might try to manipulate us into wanting tea, so that you will ask for tea.

As we can see, defining a task for our robot is not as easy as we first thought. Having the robot behave in a way that aligns with our interests is an area of Artificial Intelligence research called AI Alignment Theory.

# The Stop Button Problem:

From now on I am going to assume that we have found some way to define our task so that our robot carries out our task as we want it to. However, another problem, called the Stop Button Problem arises.

Imagine our robot is walking to the kitchen to go and make some tea. During its walk to the kitchen, it encounters the pet cat lying in the middle of the hallway. The robot is sturdy enough that if it steps on the cat it will remain standing and will be able to carry on. In this case it will continue forward and step onto the cat, as it will take longer to go around the cat than over it, therefore allowing it to score 10 points for making a cup of tea in a shorter amount of time.

You don’t want this to happen as you are quite fond of your cat, so you rush over and wrestle it to the ground as it fights to try and break free and make you your cup of tea. You yank out its battery to shut it down and think to yourself “There needs to be an easier way to turn this thing off…”. Because of this you install a big red stop button on the robot’s chest that will shut it down when it is pressed.

The next time you see the robot about to step on the cat, you rush over and try to hit the button, however the robot doesn’t let you. This is because it knows that if you turn it off, it won’t be able to make the tea, and so it won’t score any points.

Because of this you need to give it some incentive to let you hit the button. You try having it earn 5 points when the button is pressed, but the same problem arises as it scores more for making tea.

You change this so that it gains the same number of points for having the button pressed as it gets for making tea. You would think that this would work but now when you turn the robot on, it immediately presses the button and shuts itself down. This is because this way it gains the same amount of points, but in a much quicker time. And so now you can see how by adding a stop button, more unexpected behaviours arise.

# Solutions to these problems:

In all these scenarios the bad outcomes have been quite tame, however it is quite easy to imagine some very bad scenarios.

The tea robot could for instance, find a computer connected to the internet and hack every bank in the world and steal all the money from them, use this money to hold the world’s governments hostage, enslave all of humanity from this position of power and turn them into one global tea producing workforce.

This is obviously incredibly dramatic and would be incredibly unlikely to happen, however it does illustrate how unintended outcomes could negatively impact the safety of people.

[3]Another example of an AGI that has been given is a robot that has to push a block from one end of a room to another, with there being obstacles in its way. If we ask the AGI to move the block from one end of the room to another it will push the block to the other side of the room whilst ploughing through the obstacles, which is bad.

One method for preventing these unwanted actions is to have the AGI try to avoid changing the state of the environment, except for what is required by the current task. This is called giving the AGI an Impact Regularizer (IR).

What the IR does is specify some initial state of the environment, like when the AGI isn’t there and therefore, the environment is completely unaffected by it. It then looks at how much its possible actions will change the environment from this state and try to minimise this difference.

Another way to do this, is instead of trying to minimise this difference in the initial and upcoming environments, is to give it a “budget” of how much it can change the environment whilst performing its task.

Using techniques such as these may help to limit the damage that AGIs can and/or will do when executed, however there are still many unsolved questions when dealing with AGIs and AI Safety. Also using the technique described above doesn’t stop the AGI from causing damage, it only limits it. We still need to find better ways of dealing with these problems, otherwise AGIs may keep on making mistakes, and eventually some AGI out there might make one tremendous mistake.

Because of the seeming unpredictability and lack of control we seem to have over these unforeseen actions, there are people that believe that AGIs may be dangerous, no matter how many precautions we put in place or how well we think we have designed the task specified.

However, there are many people that believe that by the time we can develop AGIs, we will have a better understanding, and will have developed more and better techniques for dealing with the many issues that AGIs present. This would then allow us, with a still small risk of something going wrong, use AGIs to help further the technology and resources we have at our disposal.

# Conclusion:

It is difficult to decide on a stance when it comes to the safety of AGIs.

On one hand there is the unpredictability that they can present in almost any given situation and environment. On the other is the large amount of safety mechanisms that we have designed to help combat unwanted and potentially dangerous outcomes. Weather these outcomes will be enough to keep AGIs safe for use is still up for debate.

However, it is my personal belief that, by the time we are capable of producing AGIs, we will have available to us, enough safety precautions developed, and enough research done into the area of AI Safety, for us to be confident in making AGIs with fairly low risk to the environment in which they are used. I also believe that this small risk is worth taking, because if we manage to develop safe, functioning AGIs, it would help advance many areas, such as technology and medicine, which would benefit humanity tremendously.

Computer Scientists can and are helping move towards this goal by researching AI Safety, such as developing new methods of removing undesirable outcomes. One of the best outcomes that comes to mind is if someone finds a way to develop a safety precaution that can be mathematically proved to be safe and can either completely remove, or at least minimise the number of undesirable outcomes.

# References:

[1] Terence Mills, *“Artificial General Intelligence Breakthroughs To Watch Out For In 2018”*, https://www.forbes.com/sites/forbestechcouncil/2018/06/18/artificial-general-intelligence-breakthroughs-to-watch-out-for-in-2018, June 2018

I believe that this is a quality source as Forbes is known for providing up to date and reliable articles on a wide range of technology related subjects. It was also published within the last year and so the information it provides will be up to date and will reflect our current understanding in the field of AGI research.

[2] Milos Hauskrecht, University of Pittsburgh, *“Utility Theory”*, https://people.cs.pitt.edu/~milos/courses/cs2710-Fall2004/lectures/Class20-a.pdf

I believe that this is a quality source as it is a set of lecture slides used to teach Computer Science undergraduates at the university of Pittsburgh. The author of the slides is a Professor of Computer Science at the university of Pittsburgh and is continuously producing research papers.

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané, *“Concrete Problems in AI Safety”*, July 2016

I am confident that this is a quality source as it is a paper published by multiple different authors, who all specialise in Artificial Intelligence. They also come from multiple different organisations, all well known for being at the forefront of technology research and development, including Stanford University and Google.